A Bayesian General Theory of Anthropic Reasoning

David Shulman

March 29, 2011

Abstract

A non-ad hoc, general theory of anthropic reasoning can be constructed based on Bostrom[Bos02a]'s Strong Self-Sampling Assumption (SSSA) that we should reason as if the current moment of our life were a randomly selected member of some appropriate reference class of observer-moments. We do not need to use anything other than standard conditionalization of a hypothetical prior based upon the SSSA in order to estimate probabilities. But we need to make the SSSA precise. We specify exactly what is and what is not an observer, how to choose a reference class and how to select a prior probability distribution that can be used when selecting randomly from the reference class. There are both collective Dutch Book and relative frequency arguments in favor of our rules for choosing priors and reference classes. In order to handle examples like Bostrom's[Bos02a]Lazy Adam scenario, a causal anthropic decision theory is developed.

1 Introduction

Sometimes we care not just about which possible world[1] is actual but also about our temporal location or identity within the actual world. Even if I know for sure which world is actual, if I have to make a decision, it might matter whether I am George or Bill and whether the current time is noon or midnight.

We shall refer to reasoning about identity or temporal location[2] as anthropic reasoning or as reasoning about self-locating belief. We shall primarily be concerned with anthropic reasoning in the presence of uncertainty (i.e. probabilistic

[1]There are many reasons not to be completely happy with the possible worlds formalism. For example, set theoretic difficulties might arise[Kap95, Gri84]. But the possible worlds formalism is convenient. So we shall refer to a universe of all possible worlds when we really mean a set of equivalence classes of possible worlds. This set will not include all the worlds, only the relevant worlds and we shall consider two worlds to be equivalent if the differences between them are irrelevant. (What is interesting and what is relevant might be context-dependent.) When we refer to the set of all possible worlds, we are really referring to a set of competing hypotheses exactly one of which is true.

[2]I might reason about identity or temporal location either because I want to know who I am or what time it is or because I might use my information about who I am or when I am in order to help me estimate the probability that the actual world is a member of a certain class of worlds.

anthropic reasoning). We shall use the term "observer" to refer to something that is capable of reasoning probabilistically about self-locating belief[3].

Anthropic reasoning is ubiquitous and arguably necessary in both everyday life and science[Bos02b, Sus06, Pag07] and there are many anthropic reasoning scenarios that have relatively uncontroversial analyses, but there are other scenarios where anthropic reasoning might seem to lead to paradoxical or highly counterintuitive conclusions or where it is not clear how we should estimate probabilities[Bos02a, Elg00, Les92, Car83]. We would like to develop a general, non-ad hoc theory of anthropic reasoning.

Bostrom[Bos02a] has already created a general theory of anthropic reasoning based on the SSSA (Strong Self-Sampling Assumption). According to the SSSA, we should reason as if the current moment of our existence were a randomly selected element of some suitable reference class of observer-moments. It might be simpler to apply the Self-Sampling Assumption (SSA), which tells us to reason as if we were a randomly selected observer. But observers believe different things and know different things at different times in their lives, so we need the SSSA.

In order to estimate posterior probabilities, we would like to be able to use just the SSSA and standard conditionalization[4].

We can develop a general Bayesian, non-ad hoc theory of anthropic reasoning based on the SSSA and standard conditionalization. But in order to limit our freedom to make arbitrary choices, we need to make the SSSA more precise. We will also need to reinterpret standard conditionalization so that we are not necessarily conditionalizing an actual chronologically prior distribution. We might instead be conditionalizing a hypothetical prior. This hypothetical prior might not take into account some of the knowledge that we actually do have now and actually did have in the past[Mos09].

To say that we need to make the SSSA more precise is to say that we need to know exactly what is and what is not an observer, we need to know how to choose a suitable reference class of observer-moments, and we need to know how to choose a prior for that reference class. It might be too difficult to construct a truly general theory for selecting a prior so we shall assume that we have available a general nonanthropic theory of Bayesian reasoning[5] that allows us to construct a prior probability distribution for (uncentered) possible worlds, but we still need a theory to tell us how the prior probability that is given to a certain possible world is distributed among the observer-moments in that world.

It is the goal of this paper to make the SSSA precise in such a way that it is possible to construct a general theory of nonanthropic reasoning.

The organization of the rest of this paper is as follows: In section 2.1, we

[3]A more precise definition of observer will be provided in section 7 of this paper.

[4]Various generalizations and modifications to standard conditionalization have been proposed[Tit07, Mea08b, Hal04] but since standard conditionalization seems to work fairly well for nonanthropic scenarios, we might like to also use standard conditionalization in the anthropic case.

[5]One might well doubt whether we can necessarily always separate the nonanthropic and anthropic parts of the problem of choosing a prior but it is a convenient simplifying assumption.

present our centered possible worlds formalism and introduce some notation. Many of our arguments are wagering arguments. In section 2.2, we discuss when and why we should be convinced by wagering or decision-theoretic arguments.

Before describing our theory in complete generality, we find it useful to present some troubling anthropic reasoning scenarios. It is easier to explain the general theory if we have these scenarios available to use as illustrative examples. Section 3 contains a detailed analysis of the Doomsday Argument[Car83, Les92] scenario and why anthropic reasoning can lead to apparently paradoxical or at least counter-intuitive results. Section 4 is about the Sleeping Beauty[Elg00] scenario and section 5 discusses why we need to use anthropic reasoning if we are to link cosmology with observation[Bos02b]. There is a strong analogy between the scenario of section 5 and the Sleeping Beauty and Doomsday Argument scenarios. It is difficult to see how we can accept the use of anthropic reasoning in order to link cosmology with observation and yet not apply anthropic reasoning to the scenarios of section 4 and section 3 (and thus it is difficult to see how we can avoid the conclusion of the Doomsday Argument).

Section 6 describes another simple scenario. In this scenario we know which world is the actual world, but we have to estimate how likely it is that we are Al rather than a duplicate of Al. Al and the duplicate are in the exact same subjective psychological state but it is harder than it might seem to establish that we are just as likely to be Al as Al's duplicate.

Sections 7, 8, and 9 tell us how we should choose a prior. Thus we need to choose a reference class (section 9) and there are both relative frequency and collective Dutch Book arguments in favor of choosing almost maximal reference classes. We might like to say include all observer-moments in the reference class, but we cannot do that. If i and j are two different observer-moments and what j believes or how j makes decisions is constrained by what i believes or the decisions i makes, then i and j might not truly be distinct and independent observer-moments. It is not a problem if in practice j will have often almost the same beliefs as i but it is a problem if j would have almost the same beliefs as i even if it were exposed to quite different evidence or if it were impossible to expose i and j to very different evidence. We should want reference classes to be maximal classes of approximately independent observer-moments.

Section 7 explains why an observer-moment is just something that can sensibly be modelled as being capable of probabilistic anthropic reasoning. This modelling might involve a limited amount of idealization. It also needs to be kept in mind that an observer might be capable of reasoning about certain issues (certain universes of possible worlds) and not others and therefore something might be considered an observer with respect to certain issues but not others.

Section 8 analyzes how we might put a prior probability measure on a reference class of observer-moments. We restrict attention to the case where the reference class is finite and show why we might want to let the prior probabilities for the observer-moments in any given possible world be proportional to the amount of information that the observer-moments are capable of representing.

In section 10, there is a short description of how the case where the reference class is infinite might be handled by expressing infinite scenarios as limit of finite

scenarios. The problem is it is not clear how that should be done. We might have to learn from experience how to handle the infinite case. In fact, since the general theory of anthropic reasoning presented in this paper is not (even in the finite case) the only possible plausible theory and our arguments in favor of our theory are not incontrovertible, we might have to learn from experience which theory of anthropic reasoning is best to apply. But there is no question of not applying some theory of anthropic reasoning. Learning from experience which theory to apply is just standard application of Bayes' rule (section 11) or standard empirical Bayes[CL00].

Section 12 analyzes why we can often arrive at correct results even if we do not distinguish between the different observer-moments belonging to the same observer or even distinguish between the different observers in a given possible world. Thus section 12 tells us why many probabilistic reasoning problems do not require us to use anthropic methods and even problems that do require anthropic reasoning need not require that we distinguish between different observer-moments belonging to the same observer. This section also discusses why in problems like Sleeping Beauty we often do not care about irrelevant fine details of the subjective psychological states of observers.

Section 13 contains a sketch of an anthropic causal decision theory and section 14 presents a conclusion and suggestions for future research.

2 Formalism and Notation

2.1 Centered Possible Worlds

We use the concept of centered possible world[Lew79]. A centered world is an ordered pair consisting of a possible world and a center. A center might be thought of as a perspective from which we are viewing a world. If c is something that exists in the world w, then (w, c) is a centered possible world. But we are primarily interested in the case where the center c is an observer-moment.

In much of our discussion an observer-moment will be an ordered pair consisting of an observer o and a time interval t[6] during which o is an observer. But if one has to split an observer into parts, it might not be only temporal parts that are of interest to us. We might view a human observer as consisting of several different subpersonalities with different beliefs and desires. So we might consider each subpersonality as a separate observer-moment or consider observer-moments consisting of a subpersonality during a certain time-interval.

We want to refer to THE universe of possible worlds[7] but we cannot assume that every observer-moment z[8] is working with the same universe. Our decision-theoretic arguments and wagering arguments (including collective Dutch Book

[6] That time interval might just be a single point in time.

[7] actually the set of interesting equivalence classes of possible worlds

[8] Here we refer to observer-moment z but what is really meant is centered-world z. Thus we really are referring to a particular observer-moment in a particular world, but it is more natural and more convenient to talk about the beliefs and decisions of observer-moments rather than of centered-worlds.

arguments) and relative frequency arguments make most sense if we can reasonably model our observer-moments as being perfectly rational or at least as being capable of being perfectly rational. There is some idealization involved in modelling observer-moments as being perfectly rational but we do not wish to idealize too much. We might imagine a certain observer-moment as being capable of having coherent beliefs about some issues and not other issues and such an observer would be able to construct a coherent and reasonable probability distribution for certain universes of possible worlds and not others.

So for each observer-moment z, there is a universe W_z of possible worlds. As far as z is concerned, there might exist possible worlds outside of W_z but either z thinks they have zero probability of being actual[9] or z is incapable of reasoning coherently about these other worlds so that if $w \in W_z$, z will not really tells us her probability estimate for w being actual, only the conditional probabilty of w being actual given that the actual world lies in W_z. If $w \in W_z$, w is really an equivalence class of possible worlds; either z does not care about the differences between the different worlds in the equivalence class w or z has difficulty reasoning coherently about the differences.

In much of our discussion we will refer to W rather than W_z because we will only be concerned with a single universe, but in general different observer-moments can work with different universes. So we let W be a universe of possible worlds of interest to us and let $W*$ be the set of all centered possible worlds of interest to us. We assume that each $z \in W*$ is a world of the form (w, c) with $w \in W$ and that for all $w \in W$, there exists a $z \in W*$ of the form (w, c). In general if $A \subseteq W$, we will use $A*$ to represent the set of $z \in W*$ such that there exists $w \in A$ with $z = (w, c)$ for some center c. Thus $A*$ is the set of centered worlds in $W*$ that are obtained by pairing a world in A with a center. If A is a singleton set $\{w\}$, we shall write $w*$ instead of $\{w\}*$.

We assume that all $z \in W*$ have available to them a theory of nonanthropic reasoning that will make it possible for them to construct a nonanthropic prior P for W. All $z \in W*$ are assumed to use the same nonanthropic prior.

To say that P is nonanthropic is to say that no anthropic information was used either directly or indirectly to obtain P[10]. If for all $z \in W*$ knows, the world w might be the actual world and $x \in w*$, z's nonanthropic prior should not take into account any information that z has and x does not have or that x has and z does not have. If any such information is taken into account, then the nonanthropic prior is not really nonanthropic. Thus x and z should use the same prior. It is convenient for us to also require that x and z use the same nonanthropic prior even if $x \in w*$ and z knows for sure that w is not actual; we

[9] But we might need to discuss some of these zero probability worlds when considering counterfactual possibilities.

[10] Many discussions in the literature of anthropic reasoning do not take sufficient account of the possibility of indirect use of anthropic information. If some observers have made more observations than other observers in the same world, they might know more than other observers about which world is actual and they might have a more sophisticated physics. If they use this more sophisticated physics to construct their nonanthropic priors, then this nonanthropic prior is not truly nonanthropic. It might actually be difficult to construct a truly nonanthropic prior.

might be interested in how z would reason if she did not know that w was not actual.

Given $z \in W*$ and $A \subset W*$, we need to say how z should compute a posterior probability estimate for A. This estimate would quantify how likely z thinks it is that she belongs to the set A.

First z needs to take her nonanthropic prior P and use P to build an anthropic prior. That requires that a reference class R_z be chosen. This is the suitable reference class of observer-moments referred to in the SSSA. Then it is necessary for z to reason as if she were a randomly chosen element of R_z. The random selection presupposes that we can define for every $A \subset R_z$, the probability that the randomly chosen element of R_z will actually belong to A. Thus we assume that z can construct a prior probability distribution P_z on R_z. This prior needs to be conditionalized on the knowledge K_z that z actually does have. So what z really needs to do is compute the conditional probability $P_z(A|K_z)$.

The probability distribution P_z must be consistent with P and that means that if $V \subset W$, then $P_z(V* \cap R_z) = P(V)$. The real problem is how should z apportion among the various $x \in w*$, the prior probability z gives to some world $w \in W$.

At this point we need to say a little more about the K_z, the knowledge states, on which we conditionalize the prior P_z. When considering what an observer-moment knows, we assume that observer-moments have available to them a complete description of every element of $W*$ and of every element of W[11]. We also assume that observer-moments are perfectly rational (at least when reasoning about $W*$) and that if they know some proposition p, then they know that they know p and if they do not know p, then they know that they do not know that p. (This assumption really only need be true if p is a proposition about where some observer-moment lies in the set $W*$. and here by knowing p, we might only mean believing that p is certain to be true[12].)

Given our assumptions and the fact that K_z will be knowledge about which world in W is actual, knowledge about z's identity, and knowledge about the current time, we can let K_z be the set of y such that for all z knows, she might be y[13].

[11] Observer-moments know everything about all the possible worlds except they might not know which world is actual and they know everything about all observers except they may not know which observer they are and what the current time is.

[12] Only if p is certain to be true can it be necessarily safe to conditionalize on p.

[13] When conditionalizing, z does has to take into account that she knows of certain observer-moments y, that she could not possibly be y. She has to take into account the set K_z. But she does not have to also take into account the set of y in $W*$ that are such that for all she might conceivably be y but is very unlikely to be y. Jeffrey conditionalization is not necessary. If she is not sure whether she is y but thinks it very unlikely that she is y or if she is not sure if she is y but thinks much more likely that she is x, then since she is not sure if she is y, she might y. Thus y also thinks it unlikely that she is y and perhaps thinks it much more likely that she is x. In order for z to take into account her knowledge that she is unlikely to be y, she cannot conditionalize on the proposition that she is not y because she might be y. It suffices to conditionalize on the information that she is an observer-moment who believes that she is unlikely to be y (and perhaps much more likely to be x).

There is more we can say about the knowledge state K_z. Let $y, z \in W*$ and $K_y \neq K_z$. The observer-moment z knows what she knows and what she does not know and she has available a complete description of what y knows and thus she knows what she knows is not the same as what y knows. So z knows she cannot be y. So $y \in K_z$ if and only if y and z are in the same knowledge state.

If $K_y = K_z$, we might say that y and z are in the same subjective psychological state. Technically, they do not actually have to be in the same state. It suffices that any difference in their states will not affect how they reason about their location within $W*$. It might seem that in the actual world, here on Earth, there do not exist two distinct observer-moments that are in the exact same subjective psychological state. So we might find it remarkable if $K_y = K_z$ and y and z both exist in the same possible world. There must be something in the local environment of y that is different from the local environment of z. So y and z will not be observering exactly the same thing. If one takes into account things like a fly crawling up a wall or a pattern of movements of tree branches while the wind blows or some other minor details, it might seem that y and z will be perceiving different things[Nea06]. But even if they do actually perceive different things, not every little thing that is perceived is actually apprehended, not everything perceived at the fringes of awareness is truly taken in. So even if y and z are in subjectively different psychological states, they might not be able to take advantage of this difference when they reason about who they are in within $W*$. If someone were to call to an observer-moment's attention something at the periphery of her awareness, she would be able to use that information. But K_z refers to the information that z actually does have and is actually able to make use of when reasoning and making rational decisions such as decisions whether or not to accept or reject a certain wager.

2.2 On Wagering and Decision-Theoretic Arguments

Many of the arguments of this paper are wagering arguments. We start (2.2.1) by discussing wagering arguments in general and why they might be convincing. Then (2.2.2), we analyze why we might want to maximize expected total or expected average utility. In (2.2.3), we formally define the kind of wagers of most interest to us.

2.2.1 Why Care About Wagers?

Observer-moments are offered a chance to accept or reject wagers. All observer-moments are assumed to want to maximize their expected monetary return and choose to accept or reject a wager accordingly. In order to compute expected return, they will use posterior probabilities that will have been computed with the help of some theory of anthropic reasoning. If the consequences of observer-moments adopting a certain theory of anthropic reasoning and using that theory to choose whether or not to accept certain offers to wager are intuitively unacceptable, then perhaps there is something wrong with that theory especially

if there would be no such undesirable consequences if some other theory of anthropic reasoning were applied[Hit04, DP08, Bri10].

Wagering arguments are vivid but they have many problems[BL06]: If a bookie offers to bet with you, how can you be sure that she will actually pay you if you win? How do you know that the mere fact that you were or were not offered a certain wager is not useful information that should influence your decision whether or not to accept the wager; maybe the bookie knows more than you do and will take advantage of that knowledge to offer you a wager that seems very appealing but really is not? Why should one try to maximize expected monetary return? Maybe the thrill or anxiety associated with uncertainty matters. Utility is not necessary a linear function of money. Additional problems for our arguments arise if the bookie herself is considered to be an observer (rather than some kind of automaton); we do not want observer-moments to modify their probability estimates to take into account the fact that they know that they are not the bookie.

Some of these problems vanish if we view wagering arguments as really being arguments about decisions and maximizing expected utility. But the one advantage money has over units of utility is that money is transferable and it makes sense to add up the monetary returns of George and Bill[14]. It is not so obvious it makes sense to add their utilities.

It might makes sense to add the utilities of Bill and George. The real problem is that utilities are only defined up to a positive linear transformation[vNM44]. So if you multiply George or Bill's utility functions by 3, the decisions he will make will not change. So how do you know how to scale Bill's and George's utility function so that it makes sense to add their utilities? You can ask yourself this question: "If in some actual or hypothetical world I did not know whether I was Bill or George but believed myself as likely to be Bill as George and I were offered a chance to perform an action resulting in a net gain of one unit of utility for Bill and a net loss of one unit for George, would I perform the action?" If the answer is yes, then one unit of Bill's utility is worth at least as much as one unit of George's. Additional questions can be asked to help us determine the exact conversion factor between Bill utility and George utility[Har55]. Our wagering arguments do not really require that we actually solve the problem of comparing utilities of different observer-moments. They just require that a solution exists.

In practice, if George really did not know whether he was George or Bill, he would probably. have to be able to compare the utilities of Bill and George in order to know how to make decisions. More generally, we might have $x, y \in K_z$ and $x \neq y$ and z might have to make a decision that has different effects on the utilities of x and y. So z should be able to compare the utilities of x and y. Even if George knows he is not Bill and that the current time is noon and

[14]However, we might still wonder why we care about maximizing the total expected return of Bill and George. George might only care about George and Bill might only care about Bill. Who is to say that the best thing to do is to maximize the expected return of Bill plus the expected return of George rather than the expected return of Bill plus twice the return of George. Maybe George needs money more or belongs to a higher caste.

not midnight, he might be able to imagine being ignorant of what time it is and whether he is George or Bill. If he truly can imagine this ignorance, he should be able to imagine making a decision which affects differently his utility depending on what time it is and whether he is Bill or George. He will need to make tradeoffs and compare utilities to know how best to act. So we will accept the legitimacy of adding the utilities of different observer-moments.

Thus it might make sense to analyze the expected total return (summed up over all observer-moments) or expected average return that observer-moments obtain in a wagering scenario. The question still needs to be asked: Should we care about maximizing expected total utility and should we care about maximizing average expected utility?

2.2.2 Should We Care about Total Utility, Average Utility, or Neither?

Now that we have seen that we can meaningfully add utilities, the question arises should we try to maximized expected total or expected average utilities. The utilities in question might be epistemic utilities if one imagines observer-moments as being in the position of choosing a posterior probability distribution. Observer-moments want to maximize accuracy; for any given $z \in W*$, the greater the posterior probability z grants to the proposition she is observer-moment z, the greater the accuracy of her probability estimate. There is a standard way to measure accuracy for a single observer-moment using the Brier score[KM05]. But it does not matter how we measure inaccuracy for a single observer-moment as long as there is a well-motivated measure that lets us attach a utility value to a degree of epistemic inaccuracy.

But however we obtain the utility values for single observer-moments, we still need to know how to combine the utilities of different observer-moments. We will come to different conclusions as to how certain anthropic reasoning scenarios should be analyzed depending on whether we add up or average the utilities of the observer-moments in each world[KM05].

First consider why we might want to average utilities. The basic reason is that we believe the SSSA. If we are some observer-moment $z \in W*$ and are really reasoning as if we were an observer-moment randomly selected from the reference R_z and $R_z = W*$ and we do not take into account any knowledge we actually have of our identity (we do not take into account K_z), then we would want to maximize $\sum_{y \in W*} P_z(y)U(y)$ where $U(y)$ represents the utility of observer-moment y and $P_z(y)$ is the probability that when selecting randomly we would select y. Assume all utilities are finite and that the set $W*$ is finite so that we do not have to worry about infinities. Then $\sum_{y \in W*} P_z(y)U(y) = \sum_{w \in W} P(w) \sum_{y \in w*} P_z(y|w*)U(y)$ and here $P_z(y|w*)U(y)$ is a weighted average of the utilities of the observer-moments in w. We are thus trying to maximize the expected value (if we pick a world w at random) of a weighted average of utilities. But we were wondering why we should try to maximize the expected value of an ordinary, unweighted average.

If we could show that $P_z(y)$ should have the same value for all observer-

moments y who live in the same possible world, then we would be done. But we might doubt that observer-moments (or observers) in the same possible-world should be equally likely to be chosen by our random selection process. We want to use wagering or decision theoretic arguments including epistemic utility arguments to help us choose how to distribute the probability $P(w)$ among the observer-moments in $w*$. So when analyzing wagers we cannot necessarily simply assume that we should try to optimize expected average utility.

We might also want to prefer theories that when believed in by all observers result in maximizing $\sum_{w \in W} P(w) \sum_{y \in w*} U(y)$. Since we are assuming the utility functions of the different observer-moments have been properly scaled, we should believe that someone who cares equally about the utility of all observer-moments in the actual world would want to maximize the expected total utility[15]. If instead of utility, $U(z)$ for $z \in W*$ measures monetary return and all observer-moments want to maximize total monetary return (something not implausible to an economist who is concerned about total real gross national product), then maximizing $\sum_{w \in W} P(w) \sum_{y \in w*} U(y)$ makes sense. There are many plausible assumptions under which all observer-moments would prefer that all observer-moments use theory A rather than B if A results in greater expected total return.

The problem is that decisions are not made by a committee of all observer-moments in a given possible world[16]. They are made individually by each observer-moment. Every observer-moment might have to make a choice between theory of anthropic reasoning A and theory B and it might be better[17] if everyone used theory A rather than everyone using theory B but it might be still better if everyone except one observer-moment in each possible world (and it does not matter which observer-moment) used theory A and that one observer-moment used theory B.

We might view choosing a theory of anthropic reasoning (i.e. choosing the reference class and a probability distribution to use to guide the selection of a random element from the class) as a game. In this game, every observer-moment might have the goal of maximizing expected total utility or expected average utility or there might be some other common goal. But we require that

[15]If there are two observer-moment in the world who each have two units of utility, it hardly damages the other two observer-moments if their utilities are kept constant but another observer-moment is created with only one unit of utility. So an impartial benefactor would prefer to maximize the expected sum not average. This might be controversial but it is at least a reasonable possibility.

[16]Nor are decisions made by a committee consisting of all the observer-moments belonging to a given observer. Yes, George at eleven might make a decision to spend all his money and that might prevent George at twelve from gambling at all or George at eleven might find a way to force George at twelve to make decisions in a certain way, but if that happen, then George at twelve is not really making certain decisions; George at eleven is making the decisions. Especially if we are dealing with a scenario involving infinities, it might useful for some observer-moments to bind other observer-moments[AEH04] (force them to make decisions in a certain way). But each observer-moment that is actually making a decisions is making her own decision.

[17]Better meaning greater expected total utility or greater expected total monetary return summed up over all observer-moments.

every solution to this game be a Nash equilibrium. If our solution says that all observer-moments should choose A, our solution is not acceptable if any observer-moment finds it rational to defect from the solution.

2.2.3 Formalizing Wagers

It is now necessary to formally specify the kind of wagers of interest to us.

Let W be the universe of possible worlds of interest to us and let $W*$ be the set of centered worlds of interest to us. Then by definition a wager is a map f from $W*$ to $\mathcal{R}$, the real numbers. If $z \in W*$, then $f(z)$ is the utility gained by z when she accepts rather than rejects the wager. We are assuming that the amount gained by z by accepting rather than rejecting is independent of what the other observer-moments do and independent of any wagers other than f that might be offered. We can deal with a wager f that is really offered to only a few of the observer-moments by setting $f(z) = 0$ for any z who is not really offered a chance to wager. So there is no loss of generality in assuming the domain of f is all $W*$.

All observer-moments have the same information about the wager f and they are all offered the same choice to accept or reject the wager. Thus no information can be obtained from the fact that one is offered a certain wager. The same wager is offered in all worlds $w \in W$ to all $w*$.

Wagers are only one kind of decision-theoretic problem, but they are general enough for most of our purposes[18]. A wager is really just a choice between two actions, a choice which is offered (at least in principles) to all observer-moments. The choice need not be a choice with material rewards. It could be a choice between two (probabilistic) beliefs. Here observer-moments would be trying to maximize expected epistemic utility.

Wagers should not be thought of as necessarily taking place in any world $w \in W$. In fact, all worlds in W might be such that certain decision problems do not arise. And even if a certain decision problem did arise in the world w, observer-moments are perfectly rational and do not make irrational choices, but we might want to discuss what would happen if they did make an irrational choice. So we need to realize that wagers might take place in worlds outside of W.

.

3 The Doomsday Argument Scenario

In this section, we discuss in detail a simple version of the Doomsday Argument scenario[Car83, Les92] and its unfortunate conclusion (section 3.1). It appears that if we believe the Doomsday Argument we will make bad decisions. It might seem that we can avoid the conclusion of the Doomsday Argument if we apply the Self-Indication Assumption (SIA) (section 3.2), but the SIA is

[18]They are not general enough to deal with the case where a decision affects the number of observers who exist in the world.

not adequately motivated and can also lead to strange conclusions. Another approach is to use minimal reference classes (section 3.3) so that $R_z = K_z$ for all $z \in W*$ but that would allow us to be Dutch Booked and does not allow us to make proper use of anthropic information we really do have.

In a simple version of the Doomsday Argument scenari there are only two possible worlds, Doom Soon and Doom Later. These are very similar possible worlds and if we did not take into account anthropic considerations, we would have no reason to think one of these two worlds more likely than the other world. We might think that we have a good nonanthropic justification for giving Doom Soon a prior probability of 0.5.

The two worlds, Doom Soon and Doom World, might be equally likely a priori to be actual and also be very similar but there is one big difference between them. Only M observers exist in Doom Soon but $N > M > 0$ observers exist in Doom Later. Here by exist is meant exist now, have existed in the past or will exist in the future.

In order to specify an observer, we can mention her birth rank. An observer o is said to have rank i if the number of observers born before o is exactly $i-1$. An observer is said to be born at the instant it first becomes an observer[19]

We shall make some idealizations and model all observers in both possible worlds as consisting of exactly two observer-moments. If o_1 and o_2 are two different observers in the same possible world, the first moment of o_1 might live during a very different time-interval than the first moment of o_2, but if o is an observer in a possible world w, it is convenient to use the notation $(w, o, 0)$ to represent the first moment of observer o in world w and $(w, o, 1)$ will represent the second moment.

All observers begin life totally ignorant of which observer they are and which world they live in. Thus if F represents the set of all observer-moments that are the first moment of some observer, we have $K_z = F$ for any $z \in F$. But if $z \notin F$, then z knows her exact birth rank i. If $i > M$, then z is high rank and z knows the actual world is Doom Later. The question is what should z believe about the likelihood of Doom Soon if $i \leq M$ and z is low rank.

We assume that the relevant reference class will be the class of all observer-moments. We assume that there exists a constant k such that $P_z(w, (o, 0)) = kP_z(w, (o, 1))$ for any $z \in W*$ and any $w \in W$ (i.e. the ratio between the prior probabilities of $(w, (o, 0))$ and $(w, (o, 1))$ is the same for all observers and worlds)[20]. For most of our analysis there will be no loss of generality in assuming that $k = 1$. What matters is that merely learning whether or not we are an observer-moment who knows her birth rank gives us no information about whether Doom Soon or Doom Later is actual.[21] We also assume that for any

[19] We assume there are no ties or ambiguities. So there cannot be two different observers in the same world with the same birth rank.

[20] Notice that if $k = 0$, then we can, for most purposes, ignore observer-moments who do not know their birth rank.

[21] If we know that we know our birth rank, then we know our birth rank and this knowledge will be greatly relevant to estimating the probability of Doom Soon but the mere fact that we know that we know our birth rank tells us nothing about whether Doom Soon is actual.

given world w and any two observers o_1 and o_2 in w, there is no reason to give more prior probability to o_1 than to o_2.

Let z be an observer-moment who knows she is low rank (let that rank equal j) and wants to estimate $P_z(\text{Doom Soon}|K_z)$, the probability that she is an observer-moment in Doom Soon. $P_z(\text{Doom Soon}|K_z) = \frac{P_z(\text{Doom Soon}\cap K_z)}{P_z(K_z)}$ but $P_z(K_z)$ is the prior probability of having rank j and knowing that one has rank j. And that probability is $.5(\frac{1}{k+1})(\frac{1}{M}) + .5(\frac{1}{k+1})(\frac{1}{N})$ since the probability that a random observer in Doom Soon has rank j is $\frac{1}{M}$ and in Doom Later the corresponding probability is $\frac{1}{N}$ while in either Doom Soon or Doom Later the probability of a random observer-moment of rank j knowing her birth rank is $\frac{1}{k+1}$ and the probability of Doom Soon (and of Doom Later) is .5.

The probability $P_z(\text{Doom Soon} \cap K_z)$ is the probability that a random observer moment lives in Doom Soon and knows she has rank j and that probability is just $.5(\frac{1}{k+1})(\frac{1}{M})$. Thus the posterior probability of Doom Soon is $\frac{\frac{1}{M}}{\frac{1}{M}+\frac{1}{N}} = \frac{N}{M+N}$. So the posterior probability of Doom Later is $\frac{M}{M+N}$. If $\frac{N}{M}$ is very large, then Doom Soon would seem to be nearly certain to be actual. This conclusion that Doom Soon is nearly certain is highly problematic for several reasons. But the basic reason is that $\frac{N}{M+N}$ is very different from .5. Thus the probability computed using anthropic reasoning is very different from the nonanthropic probability of .5. And it would clearly be very wrong for low rank observers to claim that the true nonanthropic probability of Doom Soon is $\frac{N}{M+N}$. (And then be able to use anthropic reasoning to obtain a posterior probability for Doom Soon of $\frac{N^2}{(M+N)^2}$!)

3.1 The Doomsday Argument Conclusion

To vividly see what is problematic about the Doomsday Argument conclusion, we might consider a Neanderthal of the year 40,000 B.C.E. with birth rank 40,000,000[22] wondering about the future of intelligent life on earth. In one possible world (Doom Soon), intelligent life will soon become extinct (so we might set M equal to a number in the tens of millions) and in the other world (Doom Later), intelligent life will continue for several more tens of thousands of years at least (so we might set N equal to a number in the tens of billions). So, a Neanderthal could use a Doomsday Argument to show that Doom Soon is very likely[23].

The point is not that the Neanderthal would be wrong. Had Doom Soon been actual, the Neanderthal would have been right. The point is rather that in both Doom Soon and Doom Later the observer of rank 40,000,000 would estimate the probability of Doom Soon to be close to 1. Thus there is a .5

[22] For the Doomsday Argument conclusion to be derived, the Neanderthal does not have to know her exact birth rank; it suffices to know her approximate rank and thus to know that she is low rank.

[23] Of course, it would be unrealistic of the Neanderthal to only consider these two possibilities but a more sophisticated scenario would still lead to a rather strong tendency to believe that imminent doom is quite likely.

prior probability that both the actual world is Doom Later and the observer with rank 40,000,000 would estimate that Doom Soon is virtually certain. This seems like an extreme failure of calibration.

This failure of calibration might lead to bad decision-making: Neanderthals would think themselves justified to ignore the effects of their actions on the welfare of any people who might be alive in the year 2011; in the almost certain Doom Soon world, there is no intelligent life in the year 2011. Or we might imagine a betting scenario. In both Doom Soon and Doom Later, the Neanderthal with rank 40,000,000 would be willing to make a bet that results in a gain of 1,000 dollars (or Euros or rather units of utility because money had not been invented yet) if Doom Soon is actual but that results in a loss of 100,000 in Doom Later. This does not appear to be a wise betting strategy.

3.2 The SIA

One way to avoid the conclusion of the Doomsday Argument is to not use a prior probability of .5 for Doom Soon when calculating the posterior probability of Doom Soon. If we set $P_z(\text{Doom Soon}) = \frac{M}{M+N}$, then the posterior probability of Doom Soon will be .5. We might use the SIA (Self-Indication Assumption)[Die92, Olu02] to justify the value of $\frac{M}{M+N}$ for the prior probability. According to the SIA, we should other things being equal consider more likely possible worlds in which there exist more observers or in which there exist more observer-moments. We are assuming we have available to us some reasonable way to count the number of observer-moments in a world w. We might just add up the lifespans of all the observers in world w or we might have some other simple and natural way of counting moments.

It is difficult to know how to interpret or justify the SIA. We start with a totally nonanthropic prior and that prior has to be modified in order to take into account the mere fact that we exist (this is one way the SIA has been explained) but this modification is not taking into account any specific anthropic knowledge we have about our identity or temporal location, not taking into account any knowledge that we have and that other observer-moments do not have[24]

[24]Purely formally, we might use an indifference principle to justify a kind of SIA. A bare (possible) worlds indifference principle would say that all possible worlds have equal prior probability of being actual. This principle is rather implausible but if we really have very little nonanthropic knowledge, we might use such an indifference principle. The centered possible worlds version of the indifference principle would say that P_z should give equal prior probability to all elements of $W*$. The effect when applied to the Doomsday Argument scenario is the same as using the SIA.

But the bare worlds indifference principle is implausible. We might instead use a prior probability that depends on how simple the worlds are (i.e. the length of a description of the worlds, the Kolmogorov complexity of the worlds)[Sol64]. But if we assume that Doom Soon and Doom Later are equally complex worlds, then if N is much greater than M and Doom Later is actual, it will take many more bits to describe who we are among the observer-moments in Doom Later, then it would to describe who we are among the observer-moments of Doom Soon if Doom Soon were actual and we needed only to identify which of M observer-moments we are and thus we do not derive a prior that is equivalent to applying the SIA. Instead the prior probabilities of Doom Soon and Doom Later will be roughly equal.

Purely formally it is easy enough to precisify the SIA in the case where there is a only a finite number of observers in each possible world and only a finite number of worlds with observers. We distinguish between a preprior Q that somehow does not take into account even the fact that we exist and a prior P that is the result of modifying Q in accord with the SIA. (We are using the notation P and Q rather than P_z and Q_z because all observer-moments of interest are using the same prior and preprior.) We can without loss of generality assume that all possible worlds have at least one observer; it is simple enough to conditionalize Q to take into account the fact that observers exist in the actual world. This conditionalization will not affect the ratio of probabilities of two possible worlds both of which have observers. Somehow the SIA needs to modify that ratio in order to give more probability to worlds with more observer-moments (or more observers).

We can use the formula $P(v) = \frac{N_v Q(v)}{\sum_{w \in W} N_w Q(w)}$. Here for any $w \in W$, N_w represents the number of observers (or observer-moments) in w. In the case where N_w represents number of observer-moments, we assume that we have available a reasonable measure or count of the number of observer-moments in each world and that according to this count, the number of observers is finite in each world.

But even if the SIA allows us to avoid the Doomsday Argument conclusion, one might wonder if it is justified. One problem is how do we know that when constructing Q, we did not already take sufficient account of the number of observers and observer-moments in each possible world. Q was generated using a procedure of some sort for distributing prior probability among the different possible worlds and that procedure should take into account a complete description of each possible world and that description will include a specification of the number of observers in each world.

However, there are fairly convincing wagering arguments in favor of the SIA. If there are one million times as many observer-moments in Doom Later as in Doom Soon, we might consider a wager f such that $f(z) = 0$ if z knows her birth rank, $f(z) = 10$ if Doom Later is actual and z does not know her birth rank and $f(z) = -1000$ if Doom Soon is actual and z is ignorant of her birth rank[25] . If every observer-moment believes the SIA, then the total amount of money gained is $10N = 10,000,000M$ if Doom Later is actual, but a total of only $1000M$ is lost if Doom Soon is actual. Since whether or not we believe the SIA, we believe the prior probability of Doom Later is at least as great as that of Doom Soon, it seems that observers should accept the wagering offer[26]. But if all observers do not believe the SIA, they will all reject the offer.

This, however, is not quite an overwhelming argument for the SIA. In our wagering scenario for everyone to accept the wager is not a Nash equilibrium

[25] [Bos07] describes a very similar wagering scenario(Beauty the High Roller) in order to demonstrate that we if believe the SSSA and do not apply the SIA, we might run into problems. His scenario is a variant of Sleeping Beauty and not Doomsday and he actually does not favor the SIA but instead seems to favor the use of minimal reference classes ($R_z = K_z$).

[26] We are assuming that all observer-moments would prefer to maximize expected total return.

solution.

Assuming that all observers agree that everyone should try to the maximize the total expected return of all the observers, we can easily agree that it is better if everyone accepts rather than everyone rejecting, but a simple calculation will show that it is even better if (in both possible worlds) everyone except one observer-moment accepts rather than everyone accepting[27]. If every other observer-moment except me has already made their decision whether or not to accept and I am an observer-moment who does not know her birth rank, then I can think, "with probability .5, I am in Doom Soon and by accepting, I will decrease total return by 1000 dollars and with probability .5, I am in Doom Later and by accepting will increase total return by 10 dollars. So regardless of what the other observers do, I should reject".[28]

Another problem with the SIA is that the same argument that an advocate of the SIA could use against someone who applies the SSSA but not the SIA could be used by someone who advocates a stonger version of the SIA such as $P(v) = \frac{N_v^2 Q(v)}{\sum_{w \in W} N_w^2 Q(w)}$ against someone who advocate the usual version of the SIA with $P(v) = \frac{N_v Q(v)}{\sum_{w \in W} N_w Q(w)}$. Consider once again the wagering scenario where observers do not know their birthrank. When computing the expected total return, we used the preprior Q rather than the prior P. But if we really believed the SIA, we should use P not Q[29].

This time consider a wager that offers each observer-moment who does not know her birth rank a gain of 10 dollars if Doom Later is actual but a loss of 100 million dollars if Doom Soon is actual. Observer-moments who do know their birth-rank neither gain nor lose. If observers use the standard SIA to compute their priors, then they will reject the wager. They will believe Doom Later one million and not ten million times more likely than Doom Soon. If they believe the strong SIA, they will accept the offer. But if they all accept

[27]Then the expected total return is $.5(10N-10)-.5(1000M-1000)$ rather than $.5(10N)-.5(1000M)$.

[28]There is another wagering argument for the SIA that seems even stronger. We might once again consider the Neanderthal with rank 40,000,000 who is betting on whether Doom Soon or Doom Later is actual. This is a single observer here who does not know whether she has rank 40,000,000 in Doom Soon or rank 40,000,000 in Doom Later. In either case, she can either accept or reject the offer to wager. It seems like there is no question of several observers having to coordinate their decisions, but actually there is a problem in the coordination of actions between the observer with rank 40,000,000 in Doom Soon and the observer with rank 40,000,000 in Doom Later. We are modelling all observer-moments as being rational and having a disposition to act rationally in response to decision problems. But if an observer-moment who knew she had birth rank 40,000,000 did not know if the actual world world was Doom Soon or Doom Later but knew that if the actual world were a different world than the world it actually is, then the observer with rank 40,000,000 would act irrationally (Maybe the irrationality in question is not true irrationality but only seem irrational according to a certain controversial philosophical theory.), that would not make the observer with rank 40,000,000 in the actual world want to act irrationally also. Each element $z \in W*$ makes her own decisions.

[29]It might be claimed that probabilities and betting odds should come apart[BL06] but they should not come apart if we are rational. If we use the correct utility function U when describing the consequences of accepting a bet, then we should use our probabilities in order to determine betting odds. That is simply part of the meaning of utility and probability.

and the actual world is Doom Later, the total gain is $10,000,000M$ but there is a loss of $100,000,000M$ if Doom Later is actual. If we take seriously the SIA prior according to which Doom Later is a million times more likely than Doom Soon, it would appear better that everyone accept rather than everyone reject.

Of course, there is a similar argument that could be used by an advocate of a superstrong SIA against the strong form of the SIA. Once again we see the instability associated with the SIA.

3.2.1 The Presumptuous Philosopher

We might also doubt the SIA because of the unbelievably high probability estimates for Doom Later. These probability estimates are so high that given the fallibility of observation, it would seem that virtually no observational evidence could lead us to believe that Doom Soon is highly probable. This is essentially the Presumptuous Philosopher's Scenario of Bostrom[BC03].

In the Presumptuous Philosopher's Scenario, at a certain time in the future there are only two viable candidates for the correct theory of fundamental physics. According to one theory, the actual world is a Doom Soon world with M observer moments and according to the other the actual world is a Doom Later world with N observer moments. Based on nonanthropic considerations, we should think both theories equally likely[30] but if $\frac{N}{M}$ is huge enough (let us say the ratio is 10^{50}), then after we appy the SIA we see that the Doom Later theory is almost certainly to be preferred. This would even be true if the nonanthropic preprior probability of Doom Later were extremely small (e.g. 10^{-20}). Thus Doom Later might not actually be a respectable scientific theory. It might be a highly unlikely crank's theory.

In any case we might try to collect additional observational evidence in order to determine which theory is correct. It might even be true that there exists observational evidence that if correctly observed and interpreted would allow us to determine with certainty which theory is correct in the actual world. However, there is always some probability that any given sequence of observations is misinterpreted. The probability that we interpret a certain sequence of evidence as evidence for Doom Soon when it is really evidence for Doom Later might be small (e.g 10^{-30}) but that small probability might not be small enough to outweight the effect of the 10^{50} to 1 prior probability odds in favor of Doom Later. And those prior odds might have been not 10^{50} but $10^{10^{50}}$. In that case experimental disconfirmation of Doom Later appears to be hopeless.

It seems, however, that there is a similar problem of lack of possibility of experimental disconfirmation even if we do not apply the SIA. Consider a variation of the Doomsday Argument scenario where $\frac{N}{M} = 10^{10^{10}}$. An observer-moment

[30]It is true that it is only certain observers who know that there are only two viable candidates H_1 and H_2 for the correct theory of fundamental physics but it might still be the case that all observers know that H_1 and H_2 have equal nonanthropic probability. In the actual world because of cognitive limitations, there might actually be some observers who would not agree that H_1 and H_2 should be given equal nonanthropic probability, but we make the idealization that all observers are always perfectly ratioal.

who knows she is low rank might want to estimate the probability the actual world is Doom Soon. Her estimate will be so close to 1 that no amount of experimental evidence could cause her to think Doom Later plausible.

But a problem of lack of testability can also arise in nonanthropic scenarios. The reason for the lack of testability in the Presumptuous Philosopher scenario and in the scenario of the previous paragraph was the fact that one hypothesis had a many times greater prior probability than another hypothesis and the likelihood ratio was so great that no observational evidence could counteract the initial bias in favor of the preferred hypothesis. The probabilities in question were probabilities computed using the SIA or probabilities computed by revising the nonanthropic probabilities by conditionalizing on the anthropic knowledge we have available. But it does not matter how the probabilities were arrived at.

So consider the following nonanthropic scenario: There are two competing hypotheses: A and B. We know exactly one of these hypotheses is true. Not taking account of any anthropic information, we obtain a probability for A that is 10^{100} times the probability for B. A and B do not disagree about the number of observers or observer-moments in the actual world. For any number N, they give the same answer to the question what is the probability that the actual world has exactly N observers (or observer-moments). Although it seems virtually certain that A is true, we can try to obtain additional observational evidence to decide whether A or B is more likely. But given the possibility of misinterpreting experimental evidence, it seems that no amount of evidence could cause us to believe B more likely than A.

However, if we keep observing evidence that if treated as correctly interpreted evidence would be evidence for the truth of B, we would not simply insist on the truth of A. We might instead doubt our initial probability estimate that A is 10^{100} times as likely as B. Or we might doubt our theory about the likelihood of misinterpreting evidence. Even if B is some extremely implausible hypothesis based on an unimaginative literal interpretation of some ancient sacred text, if we keep observing evidence that appears to confirm B, we might be lead to doubt the very low prior probability that we gave to B rather than continue to doubt B.

In the case of the Presumptuous Philosopher, the untestability problem can be avoided if instead of assuming the SIA, we assume that the probability that the SIA is correct is .99999. This will bias us in favor of the hypothesis with more observer-moments but the bias will not be so enormous as not to be overridable. After all there are conflicting intuitions to deal with here. We do not have a conclusive proof of the SIA and there are counterintuitive aspects to the SIA.

More on Untestability without the SIA And if we did not use the SIA in order to analyze a Doomsday Argument scenario with $\frac{N}{M}$ very large such as $10^{10^{100}}$, we might be forced to resort to the same maneuver and say that our assumptions such as our assumptions about the nonanthropic prior or our assumptions about what the worlds Doom Soon and Doom Later are actually like, could be incorrect and that our assumptions have a probability of at least

.000001 of being incorrect. Consider the assumption that in both Doom Soon and Doom Later there is exactly one observer who eventually comes to believe she has rank 1 (and the similar assumption for any rank less than or equal to M). There are so many observers in Doom Later and even if observers usually are correct when they say they know their birth ranks, is it not possible that 1 observer out of 10^{20} makes a mistake? And that some high rank Doom Later observers mistakenly think they are rank 1 observers? Normally, we could ignore this small probability of error but when we are dealing with huge $\frac{N}{M}$, these infrequent errors have to be taken into account.

If, in fact, no observer does make a mistake in either Doom Soon or Doom Later, that is actually very strong evidence for Doom Soon since in Doom Later, there are so many more opportuninities to make a mistake than in Doom Soon. The prior probability of Doom Soon should be then even greater than .5 and it will be even more impossible to use experimental data to refute Doom Soon. But that is to be expected with fantastic assumptions. However, for the actual Doomsday Argument scenario, as we have specified it, there really is a bad problem that we cannot disconfirm the hypothesis that Doom Soon is actual.

If we want to avoid the untestability problem, perhaps we should try another approach such as the minimal reference class approach.

3.3 Using Minimal Reference Classes

In the Doomsday Argument scenario the objective nonanthropic probability of Doom Soon is .5. If we use the SIA, then the probability estimates for Doom Soon used by observer-moments of the form $(o, 0)$ will differ from the nonanthropic probability. If we apply the SSSA and use the maximal reference class $W*$, then the probability estimates of observer-moments of the form $(o, 1)$ will differ from the nonanthropic probability. One way to avoid both problems is to use minimal reference classes and not use the SIA. So we might use $R_z = K_z$ for all $z \in W*$.

If z is an observer-moment ignorant of her birth rank, then z knows nothing beyond the fact that she belongs to the reference class R_z and thus there is no relevant anthropic information that can be used to revise the nonanthropic prior probability estimate of .5 for Doom Soon. If z knows her birth rank, then again there is no additional information that can be used to modify the initial estimate. So if we use minimal reference classes, we are double-halfers. Many researcher favor the double-halfer position in this and similar scearios[Mea08a, Bos07, Coz07, Pusar]

But it would seem that as double-halfers we are not taking advantage of information that we actually do have. We start out not knowing our birth rank and estimate the likelihood of Doom Soon as .5. If later we learn that we are high rank, we are forced to conclude that Doom Later is actual. So learning that we are high rank is evidence against Doom Soon. But then learning that we are not high rank should count as evidence in favor of Doom Soon and should raise our probability estimate for Doom Soon above .5.

There is also a Dutch Book argument against being a double-halfer. If an

observer-moment totally ignorant of birth rank will believe the probability of Doom Soon to be .5, then she will believe the probability that she is low rank and Doom Soon is actual to be .5 and the probability that she is low rank and Doom Later is actual to be $.5(\frac{M}{N})$ because Doom Later has probability .5 and a fraction $\frac{M}{N}$ of the observer-moments in Doom Later who do not know their birth rank are low rank. So an observer-moment who does not know her birth rank would believe that the conditional probability of Doom Soon being actual given that she is low rank is equal to $\frac{.5}{.5+\frac{M}{2N}}$ and that is not the same as the probability estimate of .5 that she would use for Doom Soon if she in fact does learn that it is low rank. This discrepency seems problematic and will allow us to create a Dutch Book.

It is easy enough to create a Dutch Book here. Define a wager f such that $f(z) = 0$ if z is an observer-moment that is part of a high rank observer, $f(z) = 1$ if z is low rank and does not know her birth rank and z lives in Doom Soon, $f(z) = -\frac{N+M}{2M}$ if z is low rank and does not know her birth rank and z lives in Doom Later, $f(z) = \frac{N+2M}{3M}$ if z knows she is low rank and z is part of the Doom Later world, and $f(z) = -\frac{N+3M}{4M}$ if z knows she is low rank z is part of the Doom Soon world.

Clearly all these offers to wager will be accepted if all observer-moments use minimal reference classes. An observer-moment who is ignorant of birth rank will compute her expected return as $.5 - .5(\frac{N+M}{2M})(\frac{M}{N}) > .5 - .5(\frac{N+N}{2M}(\frac{M}{N})) = 0$. A observer-moment who is aware of being low rank will compute an expected value of $.5(\frac{N+2M}{3M} - \frac{N+3M}{4M})$ and $\frac{N+2M}{3M} = \frac{N+\frac{N}{3}+\frac{8M}{3}}{4M} > \frac{N+\frac{M}{3}+\frac{8M}{3}}{4M} = \frac{N+3M}{4M}$. But assuming all these offers to wager are accepted, we can see that if the actual world is Doom Later, high rank observers neither lose nor gain while low rank observers will experience a total return of $\frac{N+2M}{3M} - \frac{N+M}{2M} < 0$. If the actual world is Doom Soon, then each observer suffers a loss of $\frac{N+3M}{4M} - 1$ dollars[31]. If we are double-halfers, we are vulnerible to Dutch Books. If we use minimal reference classes, we can easily be Dutch-Booked.

[31]This Dutch Book Argument is really a collective Dutch Book argument because it adds up the returns of different observer-moments. Double-halfers maintain that the same observer before and after learning its birth-rank is a different observer-moment and there is no reason that there need be a simple relationship between the probability distributions used by distinct observer-moments. Moreover it might not be justified to add up the utilities of different observer-moments especially when they are not even in the same knowledge state; the same observer before and after learning her birth rank know different things and make independent decisions. But if we are not to be too free to be ad hoc, we need to be guided by some priniciple that says something about how the posterior probability distributions of different observer-moments, even different observer-moments in different knowledge states, should be related to each other. We need a constraint and Dutch Book arguments will provided a natural constraint. They have an advantage over other arguments that involve averaging or adding up the returns of different observer-moments in that we do not need to know anything or assume anything about the nonanthropic prior P.

3.3.1 Purely Hypothetical Priors

But if there were no time-period during which observers did not know their birth ranks, then we could not actually set up a Dutch Book. We might model each observer as consisting of exactly one observer-moment who knows her rank. But even if there is no actual Dutch Book, there is a hypothetical Dutch Book. The version of the Doomsday Argument scenario in which observer-moments always know their birth ranks and the version in which they have a stage in which they are totally ignorant of birth rank are very similar versions and should have similar analyses. It should not really matter that much whether there actually was a stage when observer-moments did not know their birth ranks.

In any case, in the one moment per observer version, there is still a problem low rank observer have of justifying a probability estimate of .5 for Doom Soon. All observers know the totally nonanthropic probability estimate for Doom Soon is .5. But this estimate is supposed to not take into account any anthropic information including the information about birth rank that all observers actually have. That estimate needs to be revised to take into account knowledge of birth rank. Evidence that one is low rank is evidence in favor of Doom Soon and thus the probability estimate of .5 needs to be increased.

To sum up, in the Doomsday Argument and similar scenarios we have a trilemma. We can accept the validity of the Doomsday Argument, we can accept the SIA, or we can accept being vulnerible to Dutch Books.

4 Sleeping Beauty

In the Doomsday Argument scenario, observers use information about their identity in order to estimate probabilities for (bare) worlds. But observers (actually observer-moments) might also use information about temporal location to help them figure out which world is most likely to be actual. Consider the Sleeping Beauty scenario[Elg00].

Sleeping Beauty is the subject of an experiment. On Sunday a fair coin is tossed. If the coin lands heads, Beauty will be given an amnesia-inducing drug only on Sunday night. If the coin lands tails, the drug will be taken on both Sunday and Monday night. Regardless of how the coin lands, the drug will not be given at any time other than one particular Sunday and one particular Monday night.

When the drug is adminstered to Beauty at night, the effect is to give her amnesia the next day. It has no effect on her rationality. Even the amnesia is only partial. She does not forget the protocol of the experiment. But the amnesia is such that she cannot tell the difference between different days during which she is under the influence of the drug. She is in the same subjective psychological state on all days when she wakes up under the influence of the drug[32].

[32]This scenario can be varied so that instead of always being the same exact state when she wakes up under the influence of the drug, she might be in a randomly selected member of a

On each day when she is under the influence of the drug as soon as she wakes up, she is asked to estimate a probability that the coin landed heads.

We might analyze the Sleeping Beauty scenario by treating each day of an observer's life as a separate observer-moment and we shall assume that for any observer-moment z and any two observer-moments y_1 and y_2 that last a single day and live in the same possible world, $P_z(y_1) = P_z(y_2)$. Then if z is an observer-moment corresponding to a day of Beauty's life while under the influence of the drug and we use $R_z = K_z$ (use the minimal reference class) and do not make the self-indication assumption, it is simple enough to see that $P_z(\text{heads}|K_z) = .5$. There is a nonanthropic probability of .5 for heads and Beauty has no anthropic information that can be used to modify this probability[Lew01]. If Beauty employed the SIA, then $Q(\text{heads}) = .5$ but after modification to give twice the probability to tails because the tails world has twice as many observer-moments in the relevant reference class, we get $P(\text{heads}) = \frac{1}{3}$ and since there is no anthropic information to condition on, we have $P_z(\text{heads}|K_z) = \frac{1}{3}$.

If instead of using the minimal, we use the maximal reference class, and do not make the SIA, then we also have $P_z(\text{heads}|K_z) = \frac{1}{3}$. If the reference class is all observer-moments, then the fact that Beauty knows she belongs to the class of moments under the influence of the drug is relevant information. Assuming that the total number of observer-days is not affected by whether the coin lands heads or tails and that all that the coin affects is whether a particular observer has amnesia on a certain Tuesday and also assuming that there is no other observer-moment that is in the same subjective psychological state as is Beauty on the days she is suffering from amnesia, then it is twice as likely in the tails world as in the head world that a random observer-day will be a day in the life of Beauty when she is suffering from amnesia. So it is reasonable to use the $\frac{1}{3}$ estimate for heads.

5 Cosmology's Link with Observation

If certain recent speculation is to be trusted, the universe is huge, so huge that anything that it is physically possible to observe will (with probability one) be observed[Bos02b]. For any set E of possible experimental evidence, there will be some observer in the universe who will observe E and that observer will actually observe E rather than just mistakenly interpret data to mean that E has been observed. That means that if w and v are any two physically possible (bare) possible worlds, then for any observer-moment o living in world w, there exists an observer-moment o' living in world v who is in the same subjective psychological state as o[33] . Thus the mere fact that E has been observed by someone cannot be taken as evidence for w being actual.

certain set S of state but knowing which member of S is her current state will not give her any information about whether the coin landed heads or tails or about whether the current day is Monday or Tuesday. A further discussion of irrelevant information is contained in section 12.

[33]Actually this is a little imprecise. It would be more correct to say that if we pick any set E of observ ational evidence and any possible world w in which there exists an observer-moment

What can be taken as evidence for w being actual is that we observe E. In v it is only freak observer-moments that observe E. In w, E is observed by observer-moments who are not freaks. In v if a random observer-moment performs the appropriate experiments and follows the appropriate observartional protocols, the observer-moment might observe E but is much more likely to observe something else; in w, a random observer-moment performing the appropriate experiments and following the appropriate observational protocols would be very likely to observe E.

It seems that in order to determine if v or w is most likely to be actual, we need to apply anthropic reasoning. For example we might apply the SSSA.

But if we allow anthropic reasoning in the Cosmology's Link with Observation scenario, it is difficult to see how we might justify avoiding anthropic reasoning and the conclusion of the Doomsday Argument in the Doomsday Argument scenario. If we modify the Doomsday Argument scenario so that instead of observers learning their rank, they only learn whether their rank is less than or equal to .9999M or not, we see that if $\frac{N}{M}$ is very large, then there is a very strong structural similarity between this scenario and the cosmology's link with observation scenario.

Let E be making the observation that one's rank is less than or equal to .9999M. In Doom Soon, all but a few freak observers make this observation. In Doom Later, only a few freak observers make this observation. Then we are essentially in the cosmology's link with observation scenario and would be applying Doomsday Argument reasoning to reach conclusions about cosmology.

6 Duplicating Al

Elga[Elg04] discusses an interesting scenario in which an observer might know which world is actual and has to determine his identity within that actual world. Elga wants to justify a certain limited indifference principle. If o_1 and o_2 are two observers[34] in the same possible world and o_1 and o_2 are in identical subjective psychological states and we know we are either o_1 or o_2, should we say that we are as likely to be o_1 as o_2? The answer yes to this question has intuitive appeal (if we are not going to say yes, then on what basis are we going to estimate relative likelihood?) based on evidential symmetry[Whi10], but perhaps there are other arguments for $P_z(x) = P_z(y)$ when x and y inhabit the same possible world and especially when $K_y = K_x = K_z$.

So we might consider a scenario in which there is only one possible world w and in w someone has created a duplicate of a certain person Al. There might have been earlier periods of his life, when Al had no duplicate but now and indeed at any time of his life after a certain time t, Al has a duplicate (Aldup). At any time when both Al and Aldup exist, they are in the same subjective

who has observed E, then if we pick a world v at random, the probability is 1 that there exists at least one observer-moment in v who will also observe E. But the imprecise version is good enough for our purposes.

[34] Elga talks about observers and not observer-moments.

psychological state. For example, right now, they are in the same state. Al (and also Aldup) wants to knows the probability that he is Al. The intuitive answer is .5.

We might try to justify this answer by comparison with some other scenarios. What if there are actually two worlds, H and T, and in both worlds a duplicate of Al is created. Whether the actual world is H (probability .1) or T (probability .9) is determined by the toss of an unfair coin. Neither Al nor Aldup knows whether they are in H or T. The coin tossing is entirely independent of the duplication and thus should not affect Al's (and Aldup's) estimate of how likely he is to be Al. But it is not clear what that probability should be so consider another scenario.

This scenario is just like the previous scenario except that in H, AlDup is killed, and in T, Al is killed. We might analyze H as containing three (relevant) observer-moments: HAL1, HAL2, HALDUP1. HAL1 and HALDUP1 are in identical subjective psychological states. HAL2 is in a different state than HAL1. T contains three (relevant) observer-moments: TAL1, TALDUP1, TALDUP2. TAL1 and TALDUP1 are in identical subjective psychological states. TALDUP2 is in a different state. Since none of the observer-moments knows which world is actual, we require that HAL1 and TAL1 (First Stage of Al in H and T worlds) are in the same subjective psychological state. We also require that HAL2 (Second Stage of Al in H world) and TALDUP2 (Second Stage of AlDup in T world) be in the same state.

HAL2 (and thus TALDUP2) need to estimate the probability that he is HAL. We might agree with Elga that the nonanthropic prior probability of H is .1, but both HAL2 and TALDUP2 have anthropic knowledge that might affect their estimate for the probability of being Al. They both know they are in same knowledge state as HAL2. If the reference class we are using consists of all six observer-moments, then we might say this anthropic knowledge is not helpful. In both H and T, one third of the observer-moments are in the same knowledge state as HAL2. Yes, but who said that P_{HAL2} gives equal prior probability to all three observer-moments in H (and also gives equal prior probability to all three observer-moments in T)? Is not that too similar to what we are trying to prove?

If they both use a minimal reference class, then both HAL2 and TALDUP2 will think that the probability he is HAL2 is .1. But that does not mean that in the scenario where Al and AlDup are not killed (so we also have TAL2 and HALDUP2 among our observer moments) that necessarily that HAL2 (and TALDUP2) should believe he is nine times as likely to be TALDUP2 as to be HAL2. Assume everyone uses minimal reference classes. That still does not tell us how the .1 total probability for H is split by P_{HAL2} between HAL2 and HALDUP2. We cannot assume equal division. That would be too similar to what we are trying to prove. Maybe HAL2 should think he is twice as likely to be HAL2 as to be HALDUP2 and twice as likely to be TAL2 as to be TALDUP2 (and then HAL2 would think he is only 4.5 times as likely to be TALDUP2 as HAL2). Since the coin flip is really irrelevant, in our first scenario where we knew which world was actual and only needed to estimate a probability that we

are AL rather than ALdup, we might well give the answer $\frac{2}{3}$. We do not have an indifference principle.

We might try another approach to establishing our limited indifference principle. Let us go back to the original scenario with no coin toss. Al and AlDup at any given time when they are both alive are in the same subjective psychological state. It would seem that both Al and AlDup would have to give the same answer to the question of what is the probability that he is Al, but would it be rational for both of them to think that probability is .75? There appears to be a simple argument for a negative answer[35].

Let f be a wager with $f(\text{Al}) = 1$[36] and $f(\text{Aldup}) = -2$ and $f(z) = 0$ for any z other than Al or ALDup. If both Al and AlDup think they are three times as likely to be Al as AlDup, they would both accept the wager because $3(1) - 1(2) > 0$. But then if they both accept the wager, the total return is $1 - 2 = -1$. Assuming both Al and AlDup agree that a greater total return is better than a smaller total return, it seems unfortunate that they would both accept the wager.

Al could reason as follows: I do not know if I am Al or AlDup. If the other guy (Al if I am AlDup and AlDup if I am Al) did not accept the offer and I also do not accept, the total return is zero. If instead I accept then since I am 3 times as likely to be Al as AlDup, the expected total return is $.75(1) - .25(2) > 0$. If the other guy accepted and I do not accept, then the other guy is 3 times as likely to be AlDup as AL so the expected return is $.25(1) - .75(2) = -1.25$ but if I accept the total return is $1 - 2 = -1 > -1.25$. So no matter what the other guy does, I should accept. Of course, it is unfortunate that the other guy could go through the same reasoning, but I am more likely to be Al.

In other words, the wagering argument is not totally convincing. But if both the solution where Al (and AlDup) think they are three times as likely to be Al as Aldup and the one where they think the two alternatives are equally likely are possible solutions to the game of choosing a probability of being Al, we might prefer the equiprobability solution since it leads to greater total return.

7 Who or What is an Observer

Our posterior probability estimate for Doom Soon in the Doomsday Argument sceario depends on who or what is considered to be an observer[37] What is an observer is context dependent. For our purposes a context is a set $W*$ of centered worlds of the form (w, c) where c is something that exist in (bare)

[35] See the analysis by [Bos02a] of the dungeon gedanken wherein the world is a dungeon with ninety blue cells and ten red cells and there is an observer in each cell who cannot see the color of her cell but needs to guess the probability that her cell is blue. The intuitive answer is .9.

[36] Assume that it is safe to use the SSA instead of SSSA or that we are only consider one moment of Al and one moment of AlDup.

[37] Of course, if z wants to estimate the probability of Doom Soon, she could use a W_z* that contains non-observers provided the nonobservers are never allowed to be part of R_z. But it is simpler just not to allow non-observers into R_z in the first place.

possible world w. An observer z (or rather an observer-moment z because an observer might have different beliefs and different abilities at different times) with respect to context $W*$ is an element of $W*$ that is capable of reasoning probabilistically about which element of $W*$ she is.

The observer-moment z knows that she belongs to K_z, but she must be capable of hypothetical reasoning about her location in $W*$. She must be able to reason about what she would believe and how she would make decisions if she did not take into account some of the knowledge she actually does have about her location in $W*$. She must be able to reason about what her probabilistic beliefs should be if, for example, for all she knew she could be any observer-moment in some set $A \subseteq W*$ with $A \neq K_z$[38].

Observer-moments with respect to $W*$ are not required to actually have precise (hypothetical) probabilistic beliefs or to have coherent beliefs about who they are among the members of the set $W*$. But it has to be the case that if they were faced with a wager about who they are among the elements of $W*$, they would be able to act as if they were formulating sufficiently precise, sufficiently coherent probabilistic beliefs and be able to maximize their expected utility based on these beliefs when deciding whether to accept or reject the wager. It might be the case that in any possible world that an observer-moment thinks has any chance of being actual and in any possible world that might have any chance of being actual if it were the case that she had less anthropic knowledge than she actually does have, she will not be faced with a decision problem that will require her to think rationally about who she is among the elements of $W*$, but she might still have the capacity for such rationality and thus might be considered an observer-moment with respect to $W*$. We need to require rationality of observers in order for our wagering arguments to make sense.

It might not be enough to assume rationality of all observer-moments in $W*$. Our wagering arguments presuppose all the observer-moments using the same nonanthropic prior P. In the actual world, even expert observer-moments might disagree about what is the correct nonanthropic prior P. But they still might be able to use and understand the motivation for the correct prior P. We assume that all observer-moments, even if they do not know the correct theory of nonanthropic reasoning, could learn that P is a reasonable prior even if not necessarily the optimal prior. Thus it might make sense to imagine all observer-moments as using P.

But we would like to say that most adult humans are observers. Yet humans have limited ability to reason coherently and exactly about probabilistic knowledge and perhaps limited ability to understand how to calculate the correct nonanthropic prior. A certain amount of idealization is inevitable. But our theory of anthropic reasoning is a normative theory, not a descriptive theory. It might be still be useful to model humans as if they had the capacity to be perfectly rational in a certain context.

Taking into account the fact that humans can use aids such as computers or

[38]Of course, any $y \notin K_z$ knows that she is not z and thus it might seem that is impossible to reason as if one might be y and one might be z, but one can reason as if one did not know anything other than $K_z = A$ and as if no element $y \in A$ knew anything other than $K_y = A$.

experts to help them make probabilistic computations and that humans might become more rational then they normally are when the stakes are sufficiently high, it might not be too much of an idealization to model a human as if she were an observer. Moreover, we are only requiring observerhood with respect to a limited context. In any case we might say that much of the human failure of rationality is a failure of performance and not competence. There might be a set S of probabalistic statements about $W*$ such that some human understands S and can solve simple problems that require the use of S as premises bur cannot solve problem that would require the application of long chains of mathematical and logical reasoning using the premises S but if the human had the capacity for handling computational complexity, then she would be able to reason coherently about which of the elements of $W*$ she is likely to be.

7.1 But which context $W*$ should be used?

An observer-moment might be an observer-moment with respect to several different contexts (i.e sets of centered worlds). So the issue arises when estimating posterior probabilities which context should an observer-moment use. The answer in general is use the largest possible context. A low rank observer-moment in the Doomsday Argument scenario might be able to use the anthropic information that she is low rank rather than high rank together with her knowledge that low rank observers are atypical in Doom Later to derive the conclusion that the actual world is rather unlikely to be Doom Later. But this requires that she use a context $W*$ that includes high rank observer-moments although she knows she is not high rank.

So, if z is an observer-moment with respect to contexts A and B with $A \subset B$ and different posterior probability estimates would be obtained using B rather than A, it is preferable to use the larger context, B.

8 Putting A Measure on Observer-Moments in the Same Possible World

We need to return to the question of choosing an anthropic prior P_z given a nonathropic prior P. We have already seen (in section 6) why we might prefer that if $x, y \in K_z$ and $x* = y*$, then $P_z(x) = P_z(y)$. But that is surely not the only reasonable choice. Maybe, we should let $P_z(x)$ depend on the length of the minimum length description of x (and not just on the subjective psychological state of x). But the case where $x, y \in K_z$ is not the only case where we need to compare $P_z(y)$ and $P_z(x)$.

8.1 What if x and y are in the same possible world but not in the same knowledge state?

We might like to generalize the equiprobability assumption to the case where $x \notin K_y$ but x and y are in the same possible world[39]. But it does not seem correct to necessarily give equal prior probability to x and y if x exists for a longer time than y or x is capable of representing more information than y. We shall eventually generalize the equiprobability assumption in a way that takes into account the amount of information that can be represented. But first we shall discuss a crude approximation that involves the concept of atomic or indecomposable atomic-moments. As crude as the concept is, in practice people often do model observers as consisting of a succession of time-slices such that no significant belief change can take place during a time-slice but radical changes can take place in the transition from one time-slice to the next. So the concept of indecomposability is worth investigating.

If an observer-moment is indecomposable, it cannot be meaningfully expressed as a union of smaller observer-moments (i.e. observer-moments with shorter durations). If an observer-moment o exists during the time-interval t_1 and also during the time interval t_2 and believes different thing during t_1 than during t_2, then the observer-moment is definitely not atomic. There might be no great difference between what is believed during t_1 and what is believed at t_2, but if there is any disagreement even if only about what time it is, then we cannot say that o has definite beliefs or a definite knowledge state and thus o is not atomic. If o is capable of reasoning about whether the current time is part of the time interval t_1 or part of the time interval t_2 or neither and has different beliefs about that issue during t_1 than t_2, than o is not atomic and o does not have a definite knowledge state. Our analyses up to this point assumed that observer-moments had definite knowledge states. That is one reason that atomic observer-moments are interesting.

Even if in fact o believed exactly the same things (including beliefs about what time it is) during t_1 and t_2, o might still not be atomic. Maybe the only reason that the same things were believed during t_1 as during t_2 is that o was exposed to the same evidence during the two time-intervals but if she had been exposed to different evidence (and she could have been exposed to different evidence) she would have had different beliefs. In that case, o is not atomic. Perhaps she does have a definite knowledge state, but she might not have had a definite knowledge state. What matters is capability. It might be true that in any world that o believes has any chance of being actual, o will not be exposed to different evidence during t_1 and t_2, but it still might be true that if she were exposed to different evidence, she could have different beliefs during t_1 and t_2. Of course, she could have different beliefs if her neurophysiology and psychological capabilities were different then they actually are, but I am concerned with what she could believe given her actual physical and psychological capabilities. If she is actually capable of believing different things at different subintervals

[39] And we also require that x, y both belong to the relevant reference class R_z for some $z \in W*$.

of the time-interval during which she exists, then o is not atomic and she is decomposable. If she is not capable, then she is atomic.

It is actually rather an extreme idealization to view an observer i as a succession of atomic observer-moments such that if i_1 and i_2 are atomic observer-moments belonging to i with i_2 occurring immediately after i_1, then no belief change is possible at all during the time interval when i_1 exists or during the time interval when i_2 exists but abrupt change is possible at the transition from i_1 to i_2. But it is a convenient crude approximation that can later be refined.

In fact, we need to refine our criterion for atomicity. If an observer-moment o exists during the time intervals t_1 and t_2, what matter is not whether o can believe different things during t_1 as during t_2 but whether her beliefs at t_2 could be independent of her beliefs at t_1 if she were exposed to (and she could be exposed) to very different evidence. If it were inevitable that what she believed at t_2 is totally predictable given what she believed at t_1 (and independent of any additional evidence she had at t_2, then regardless of how different her beliefs might be at t_2 from what they were at t_1, we could not really decompose o into o at t_1 plus o at t_2[40]. We do not have independence. We have total dependence. It would be wrong to think of the observer-submoment at t_1 and the observer-submoment at t_2 as truly different if the belief during t_2 are fixed given the beliefs during t_1.

We might also have partial dependence. In practice, o during t_2 maybe should believe something very similar to what was believed at t_1 because, for example, t_2 occurs right after t_1 and it takes time for the state of the world to change very much, but if the world did change very much and there were evidence of such change and yet o during t_2 is in part constrained from forming her beliefs solely based on the evidence available to her (including any communication (i.e. memory) she might have received from the o during t_1 observer-submoment) and forced to believe certain things solely because they were believed by o during t_1 or if it is impossible for o during t_2 to observe very different evidence than the evidence observed by o during t_1, then we might say that we have partial dependence and not independence.

In practice, we will always have some dependence. The further apart in time the two time-intervals, t_1 and t_2 are, the less dependence we expect. But there will be some dependence.

We shall, however, start with an idealized model each observer consists of a finite set of totally independent observer-moments. Each observer-moment then can say of any $y \in W*$, "yes, for all I know, I might be y" or "no, I know I am not y". Thus all observer-moments are representing the same amount of information. They all represent $|W*|$ bits of information where $|W*|$ represents the number of observer-moments in $W*$. Furthermore there is no redundancy between what one observer-moment is capable of representing and what another observer-moment is representing.

[40] Of course, instead of the observer-moment at t_2 being dependent on the observer-moment at t_1, we could have dependence in the reverse direction or in both directions and this is true regardless of whether t_1 occurred before or occurred after t_2

To say that there is no redundancy is to say that for any $V \subset W*$ and any $z \in W*$, it could be the case that z is exposed to evidence that insures that $K_z = V$ and this is true regardless of what is believed or what evidence is available to any other observer-moment. But we need to be careful about the "could be the case". It most likely will not be the case in any world that has nonzero probability according to P. But z is physically and psychologically capable of obtaining evidence that would lead her to have the knowledge state V regardless of what the other observer-moments believed, or knew and regardless of the evidence they were exposed to.

In this idealized model of independent atomic-moments, it is appealing to say that $P_z(x) = P_z(y)$ if $x* = y*$ even if $K_x \neq K_y$. After all, even if x and y are not in the same knowledge state, they could have been. And if they were, they would have the same prior probability. Symmetry considerations should tempt one to think that $P_z(x)$ should not depend on exactly what x does believe, but on what it could believe. It is difficult to see how we could justify a particular rule for making $P_z(x)$ depend on K_x.

But if we do assume that $P_z(x) = P_z(y)$ if $x* = y*$, then for any $A \subset x*$, $P_z(A)$ is proportional to the total amount of information that can be represented by the observer-moments in A. It is tempting to generalize this information-theoretic criterion to the case where $W*$ is not just a set of independent observer-moments. So even if independence and atomicity assumptions are false of $W*$, we would say that if $x* = y*$, $z \in W*$, $A, B \subseteq x*$ then $\frac{P_z(A)}{P_z(B)}$ is equal to the ratio of the total amount of information that can be represented by the observer-moments in A to the total amount of information that can be represented by the observer-moments in B. By total amount of information that can be represented by A or B, we mean the number of bits needed to represent the knowledge states of all the observer-moments in A or all the observer-moments in B.

This information-theoretic criterion is useful, but in most of the rest of this paper, we idealize and assume that either $W*$ is a finite set of independent atomic moments or that, in any case, we can reason as if $W*$ consisted of a finite number of independent atomic moments and that means that for any $z \in W*$, any $w \in W$, any $A \subseteq w*$, $P_z(A|w*) = \frac{|A|}{|w*|}$. Here $|\quad|$ means cardinality of. Thus the prior probability of A given $w*$ is equal to the number of observer-moments in A divided by the total number of observer-moments in world w.

Now that we have a simple way of splitting the prior probability of a world among the observer-moments that belong to that world and the relevant reference class, we need to discuss how to choose that reference class.

9 Choosing A Reference Class

The posterior probability estimates $P_z(A|K_z)$ depend on the reference classes R_z. We knows that $K_z \subseteq R_z$. If $y \in K_z$, then z does not know that she is not y. So $P_z(y|K_z)$ must be meaningful and thus $y \in R_z$. But if we choose $R_z = K_z$, we will have trouble linking cosmology with observation (see section 5). In this section we discuss how to choose a reference class under the assumption that

all the finitely many observer-moments in $W*$ are independent and that for any world w, if $x, y \in w$ and $z \in W*$, then $P_z(x) = P_z(y)$. Under these assumptions we can provide both collective Dutch Book and relative frequency arguments in favor of using the maximal reference class $R_z = W*$. If we can prove that we would obtain the same posterior probability estimate $P_z(A)$ if we used a smaller reference class, then we might use a smaller reference class, but, in principle, we should use maximal reference classes.

9.1 A Collective Dutch Book Argument

Our collective Dutch book arguments involve showing that under certain assumptions if observer-moments use non-maximal reference classes, then a collective Dutch Book can be constructed. That means there exists a wager f that could be accepted by all observer-moments[41] such that if observer-moments use non-maximal reference classes when calculating their posterior probabilities and observer-moments make decisions in a such a way as to optimize their expected utility, then in no possible world w is the sum $\sum_{z \in w*} f(z)$ positive and in some worlds, it is negative. But that means that regardless of which nonathropic prior P we use, if we were to pick an observer-moment y at random by first using P to pick a world w at random and then giving each $z \in w*$ an equal chance of being selected, the expected value of $f(y)$ would be negative.

We can generalize the Dutch Book we constructed when discussing the Double Halfer analysis of the Doomsday Argument (and Sleeping Beauty) scenarios. Assume that there exists at least two different possible worlds v and w and four different observer-moments $v_1, v_2 \in v*$ and $w_1, w_2 \in w*$ with $K_{v_1} = K_{w_1}$ and $K_{v_2} = K_{w_2} \neq K_{w_1}$. Thus all we are really assuming is that there are two different observer-moments in different knowledge states who do not know whether the actual world is v or w. If observer-moments living in the same possible world have the same prior probability and if all four observer-moments use maximal reference classes and c_2 is defined so that $P_{w_2}(w_2|K_{w_2}) = c_2 P_{w_2}(v_2|K_{w_2})$ and c_1 is defined so that $P_{w_1}(w_1|K_{v_1}) = c_1 P_{w_1}(v_1|K_{v_1})$, we should have $c_2 = c_1$ Thus if w_1 thinks she is c_1 times more likely to be w_1 than v_1, the observer-moment w_2 will also think that she is $c_2 = c_1$ times more likely to be w_2 than v_2. If $c_1 > c_2$, then a Dutch Book can be constructed.

Without loss of generality, we might assume that $c_1 > 1$. Then choose two numbers $a > 1$ and $b > 1$ such that $c_1 > b > a > c_2$. Let f be a wager such that $f(w_1) = 1$, $f(v_1) = -b$, $f(w_2) = -1 - \epsilon$, $f(v_2) = a(1 + \epsilon)$ and for every other observer-moment z, $f(z) = 0$. All observer-moments would be willing to accept this wager. The observer-moments v_1 and w_1 think it c_1 times as likely that they win 1 as that they lose $b < c_1$. The observer-moments v_2 and w_2 think it c_2 times as likely that they lose $1 + \epsilon$ as that they win $a(1 + \epsilon)$ but $a > c_2$.

We do have a Dutch Book. In world w, the total return is $-\epsilon$, which is negative if $\epsilon > 0$. In world v, the total return is $-b + a(1 + \epsilon)$, which is negative

[41] I write "could be accepted" because for many observer-moments z, $f(y) = 0$ for all $y \in K_z$ and it does not matter whether such obsever-moments accept or reject. So we really require that all observer-moments believe that accepting is at least as good as rejecting

if $a(1+\epsilon) < b$ and it is certainly possible to choose a positive ϵ that makes this true.

If maximal reference classes are used, we will certainly have $c_2 = c_1$. But if w_1 uses a nonmaximal reference class R, let R_{1w} be the number of observer-moments in w that are in R and R_{1v} be the number of observer-moments in v that are in R. Then given the fact that all observer-moments in the reference class that lives in the same possible world has the same prior probability, we have that $c_1 = \frac{P(w)R_{1v}}{P(v)R_{1w}}$ where P represents the nonanthropic prior. With obvious notation we also have $c_2 = \frac{P(w)R_{2v}}{P(v)R_{2w}}$. There is no reason then that c_1 and c_2 need be equal.

Our Dutch Book result can be generalized to the case where we have worlds $w_1, w_2 \ldots, w_n$ and for $1 \leq i \leq n$ there exists observer-moments $a_i, b_i \in w_i$ such that for all $1 \leq i < n$, a_i and b_{i+1} are in the same knowledge state and also a_n and b_1 are in the same knowledge state, but each a_i is in a different knowledge state.[42]

9.2 A Relative Frequency Argument for Maximal Reference Classes

Our other argument for the use of maximal reference classes is a relative frequency argument. We shall just illustrate how the relative frequency argument would work when applied to the Doomsday Argument scenario. So assume that instead of just having a Doom Soon and a Doom Later world, we have many different dimensions in which possible worlds can vary. Thus a world w will be characterized by a binary digit number a of length n. Let a_i represent the ith digit of a. If $a_i = 0$, we say the world is Doom Soon in dimension i. If $a_i = 1$, it is Doom Later in dimension i. If v and w are two different n-digit numbers, then the actual world is as likely to be characterized by v as by w. So the prior probability of any particular world is 2^{-n}.

Observers are characterized by a multidimensional rank. Their multidimensional rank is an n digit binary number. For any $1 \leq i \leq n$, if a world w is Doom Soon in dimension i, then all observers have rank 0 (low rank) in dimension i,

[42]Let $c_i = \frac{P_{a_i}(a_i)}{P_{a_i}(b_{i+1})}$ if $1 \leq i < n$ and $c_n = \frac{P_{a_n}(a_n)}{P_{a_n}(b_1)}$. Then if we use maximal reference classes we should have $\prod c_i = 1$. If not all observer-moments a_i use maximal reference classes, we might not have $\prod c_i = 1$. If not, a collective Dutch Book can be constructed.

Without loss of generality assume that $\prod c_i > 1$. Let $0 < r < 1$ be a number that we will want to be just slightly less than one. We define a wager f so that $f(z) = 0$ except for observer-moments of the form a_i or b_i. If $1 \leq i < n$, we will have $f(a_i) = r^{2i-2} \prod_1^{i-1} c_j$ while $f(b_{i+1}) = -r^{2i-1} \prod_1^{i} c_j$. $f(a_n) = r^{2n-2}$ and $f(b_1) = -r^{2n-1} \prod_1^{n} c_i$. Then an observer-moment who thinks she is c_i times more likely to be a_i than b_{i+1} will accept the wager as will an observer-moment who thinks she is c_n times more likely to be a_n than b_1. So everyone accepts the offer. But in world w_i with $2 \leq i \leq n$, the amount gained by a_i is r times the amount lost by b_i and thus the sum of the returns is negative. In world w_1, the sum of the returns is $1 - r^{2n-1} \prod c_i$ and we can choose r close enough to 1 to make this difference negative. So we have a collective Dutch Book.

but if a world is Doom Later in dimension i, then an observer can be either low rank or high rank (rank 1) in dimension i[43].

Each observer consists of 2^n observer-moments. We might use an n binary digit number to characterize the observer-moments belonging to a given observer. For $1 \leq i \leq n$, if the ith digit of the number characterizing an observer-moment is zero, then the observer-moment does not know her rank in dimension i, but if the digit is 1, then she does.

The scenario we have constructed is a scenario involving n independent repetitions of the Doomsday Argument scenario. It is important here when we try to justify our relative frequency argument that we are really dealing with independent experiments. This scenario really does involve n truly independent repetitions of identical Doomsday Argument scenarios.[44]

Assume n is very large. Then the probability that the actual world is Doom Soon in approximately half its dimensions is very close to 1. Thus a typical world will have about as many Doom Soon as Doom Later dimensions. Now consider a typical observer in such a typical world. In about half the dimensions the world is Doom Soon and of course the observer is low rank in these dimensions. In the other half of the dimension the world is Doom Later and in those dimensions there are as many low rank as high rank observers. Thus a typical observer will be low rank in about half the dimensions in which the world is Doom Later and that means a typical observer in a typical world will be low rank in Doom Later in about one fourth of the dimensions. Hence in approximately two thirds of the dimensions in which she is low rank, the world will turn out to be Doom Soon.

Notice that if we are a typical observer-moment belonging to a typical observer in a typical world and use maximal reference classes and the SSSA and do not apply the SIA and for a dimension i in which we know our rank is low, we try to estimate the probability that the actual world is Doom Soon in dimension i, we will obtain a posterior probability estimate of $\frac{2}{3}$. So we will be well-calibrated.

[43] We could easily generalize our scenario to allow low ranks to range from 1 to M and high rank from $M+1$ to N.

[44] Repeated Sleeping Beauty scenarios discussed in the literature(for example in [Bos07, Arn02]) do not really involve independent repetitions. We might consider n different variattions of our simple Sleeping Beauty scenario in which a drug is given either once or twice. The only real difference between these variations is that a slightly different drug is administered to Beauty in each variant, a different coin is flipped and the drug is administered during different weeks. In fact n different drugs are given in n successive weeks. But then if we consider just one particular drug, any day of Beauty's life is a either the first day during which Beauty is under the influence of the drug, the second day during which she is under the influence, or a day during which she is not under the influence. But whether or not she is under the influence of one drug is not indepedent of whether or not she is under the influence of another drug. Our repeated Sleeping Beauty scenario involves different drugs, but we might imagine that the effects of these different drugs on Beauty's subjective psychological state are indistinguishable. The fact that different drugs were involved during different weeks will not affect Beauty's probability estimates when guessing what happened when coins were tossed. So we really are discussing the same Repeated Sleeping Beauty scenario that is discussed in the literature. And we do not have independence(See also [Kad04] for a discussion of the lack of independence.

We can modify our scenario so that observer-moments will even be able to learn that they will be well-calibrated if they believe the Doomsday Argument conclusion that the probability of Doom Soon is $\frac{N}{M+N}$. (In this section we are letting $N = 2$ and $M = 1$.) Let each observer in the (unrepeated) Doomsday Argument scenario consist of three not two observer-moments. The third observer-moment knows both her rank and whether the actual world is Doom Soon or Doom Later. The other two observer-moment are just as in the earlier version of the scenario. The third moment knows (remembers) any probability estimates made by the second moment of the same observer. The repeated version of this scenario is just like the repeated version described above except that moments belonging to a given observer are characterized by a ternary, not binary digit number of length n. Thus if n is large enough, a typical observer-moment will come to realize that she is well-calibrated if she guesses two-third when asked to estimate a probability for Doom Soon in a dimension in which she knows she is low rank but does not know if the world is Doom Soon or Doom Later in that Dimension.

We have discussed a relative frequency argument for a particular scenario, but we should remember that if we are to have confidence that a theory of anthropic reasoning really works, we will have to apply that theory many times to many different problems. So we might use actual rather than hypothetical relative frequencies to justify our belief in a certain approach to anthropic reasoning. It will be easier to rely on these actual relative frequencies if we are computing relative frequencies of useful predictions in independent experiments. So it is important that in our repeated Doomsday Argument scenario, we are dealing with independent dimensions.

9.3 Reference Classes are Not Too Big

It might be objected that our reference classes are too large if we use maximal reference classes as the previous section recommended. In order to calculate a posterior probability estimate for some (bare) possible world being actual, do we really have to know about the number of observer-moments in some distant galaxy, especially if intelligent life in that galaxy has a very different psychology than human psychology? The answer is, in principle, yes, we do. That is unfortunate, but that is the way it is. If we know very little about that distant galaxy, we might make some simplifying assumptions because that is the best we can do. But we make similar simplifying assumptions when using nonanthropic reasoning to evaluate cosmological theories or even to predict future events here on earth. There is often a large collection of factors that could affect the value of some variable of interest to us and some of these factors might be obscure and difficult to investigate so we make simplifying assumptions.

In any case our maximal reference classes are not all that big. They are not universal reference classes. We require that all $z \in W*$ could agree on a nonathropic prior P on W. And of course, all $z \in W*$ have to be able to actually be observer-moments with respect to the context $W*$. Perhaps we and some very alien observer-moments living in another galaxy have fundamental intuitions

that are so different that we and they could never agree on a reasonable P or we might not be able to reason coherently about the hypothetical possibility that we are just like those aliens from the other galaxy.

There is also a less speculative and fantastic way in which our maximal reference classes are not too large. We might consider whole communities or even whole civilizations as if they were observers. There is such a things as group belief, group intention, group utility[Gil89]. But our reference classes do not have to include both individual observers or observer-moments and group observers or observer-moments.

To see why we might exclude either group or individual observer-moments from our reference class, note that the beliefs and desires of a group are very much not independent of the beliefs and desires of its members. Moreover it might not be possible for either groups or individuals to reason coherently about a hypothetical situation where they do not know whether they are an individual or a group. In addition, there might be some (nonanthropic) issues about which only whole civilizations and not individual scientists can (or truly want to) reason coherently; cosmology might be so difficult for individual scientists and maybe it is not the preferred goal of the individual scientist to seek and advocate the truth as best she can; instead it might be the preferred goal that she do her part in helping the consensus of the scientific community as a whole to eventually come much closer to the truth and that, given human limitations, might require her to advocate incredible, but heuristically fruitful hypotheses, for example, and really belief these hypotheses and conduct research as if she believed these hypotheses. If on some issues only the community as a whole is rational, then on some issues we will be working with a reference class consisting just of whole civilizations rather than every single observer-moment[45]

10 The Infinite Case

Up until now we have only considered the case where $W*$ is finite. Thus the number of worlds that have observers is finite and in each such world the number of atomic moments is finite (or the total amount of information that can be represented by all the observer-moments in a given world is finite). The infinite case is much harder and it is less simple to come up with a good prior for the infinite case just based on general theoretical principles and a good nonanthropic prior.

We still want to be able to say that if i and j are atomic observer-moments belonging to the same possible world they should have equal prior probabilities but that assumption is not as useful or convenient or even convincing as it was in the finite case. Even if we accept this equal probability assumption, many probability estimates remain undetermined even given the nonanthropic prior.

But we expect infinities to create problems for us. They create problems

[45]Thus Vilenkin[Vil95]'s principle of mediocrity according to which we should reason as if we were a random civilization might make sense.

for nonanthropic scenarios[46] . The usual way of handling an infinite case is to treat it as a limit of finite cases that we can handle. The problem is that there are many ways of representing an infinite case as a limit of finite cases and not all ways result in equivalent probability estimates. We might well expect there will be no good theoretical guidelines for choosing a particular preferred way to represent the infinite as a limit of the finite. We might have to learn from experience just how to represent the infinite. The measure problem might ultimately be an empirical problem.

We can just say a few things about how we might represent an infinite case as a limit of finite approximations. If we ignore certain observers or observer-moments, our scenario might become simpler. So if we approximate an infinite case by saying that an infinity of the observer-moments should be given probability zero and only a finite number will have nonzero probability, then we have a more tractable approximation. Another way to approximate the infinite by the finite is to use equivalence classes of observer-moments and equivalence class of possible worlds. So we refuse to distinguish between possible worlds that are in fact actually distinct and refuse to distinguish between observer-moments within a given possible world that are in fact distinct. If we have only a finite number of equivalence classes of possible worlds and within each equivalence class only a finite number of equivalence classes of observer-moments then we are dealing with a finite case. Our finite approximations could also involve both equivalence classes and consider only a finite subset to be interesting.

Once we have constructed a set of finite approximations, some of which are more accurate approximations than others, we might be able to express an infinite case as a limit of an infinite series of finite approximations. The main issue is which finite approximations are most relevant and how do we express the infinite as a limit of the finite. Aside from saying that our approximation procedure and our procedure for choosing how to express the infinite as a limit of the finite approximations must seem simple and natural and lead to intuitively reasonable results, there is nothing much more that will be said about the infinite case in this paper. The key idea here anyway is that there is much that must be learned from experience.

11 Learning How to Reason Anthropically from Experience

There are many possible procedures for estimating probabilities in anthropic scenarios. Different researchers have recommended the use of different theories.

[46] Consider, for example, [Jay03]'s discussion of marginalization paradoxes. We might also think about something as simple as the need for improper priors and the problems they might cause; if we want to say that we are picking a positive integer at random and all positive integers are equally likely to be selected, we need to use an improper prior. Now I might say conditionalize on the information that the number that was selected was divisible by 3 and ask for the probability that the number chosen was even. There is no uniquely justified probability estimate here. It depends upon how we analyze this infinite scenario as a limit of finite scenarios.

Disagrement on handling anthropic scearios continues to exist even under the assumption of universal agreement about nonanthropic cases. We definitely do have to reason anthropically even if reasoning anthropically only means ignoring the anthropic information or using the minimal reference class or following some simple procedure for replacing an anthropic by a nonanthropic scenario.

We might think that anthropic reasoning has been falsified by experience[Olu04] because in fact we are not typical in a certain respect. But of course everyone and every possible world is atypical in some respect[HM05]. If we have to choose a twenty digit number at random we might pick some number s that is atypical in that it is the only number equal to s (and we might even discover a less gerrymandered way of expressing the atypicality of our having selected s). But what matters is being atypical in interesting respects. What if it turns out that $s = 11,111,111,111,111,111,111$? This s is atypical in an interesting respect in that its atypicality is defined by a property that might be likely to occur to us before the number was selected. However, it is still possible that we might seem to be atypical in some interesting respect. That would not mean that anthropic reasoning has been falsified, just that we might need to define a different formulation for anthropic reasoning and a different analysis of what it means to be atypical. Maybe if i and j belong to the same world, $P_z(i)$ and $P_z(j)$ should not be considered equal. There is still much unexplored space in the realm of theories of anthropic reasoning. And, of course, we should expect that there is someone in the actual world who actually is atypical in an interesting respect. All we can assume is that there is low prior probability of our turning out to be atypical in an interesting respect.

In any case, we know how to test theories of anthropic reasoning using standard Bayesian methodology. If X is a theory of anthropic reasoning that is sufficiently precise that given a nonanthropic prior probability distribution all necessary posterior probabilities can be calculated for relevant observer-moments, we know that although there might be strong arguments in favor of X, these arguments will not be incontrovertible and will not be based on pure logical analysis. There will be some element of intuition about naturalness, simplicity or some other doubtable assumption involved. The arguments might be very convincing but there will still be some chance that X is incorrect. Researchers might differ as to what this chance is, but they should all agree that it is meaningfully greater than 0 and less than 1.

There might be many reasonable theories of anthropic reasoning that are in contention. Certainly there are several possible approaches to handling the infinite case. We might need to choose between two competing theories X and Y of anthropic reasoning. An observer-moment would choose based on the evidence E she has available. Based on intuition, based on general nonanthropic considerations, based even on some anthropic evidence obtained prior to the acquisition of evidence E, one might have some prior probabilities for the likelihoods of X and Y. Thus we are using an actual prior knowledge state or a hypothetical prior knowledge state representing what we would know if we did not know E to estimate a prior probability for X and for Y. There might be some lack of clarity or controversy about what these prior probabilities should

be. Intuitions are not always clear. Thus we might have to work with ranges of values rather than specific prior values. Based on X or Y (our theories for computing anthropic probabilities), we can compute a probability for E given X and for E given Y. Then we can just apply Bayes' rule and hope that the evidence E will be decisive enough that regardless of where in the range of possibilities the priors for X and Y lay, one of the two competing metatheories is strongly preferred over the other.

12 When to Ignore Distinctions

In this section, we explore equivalence relations and when it is possible to correctly estimate posterior probabilities while ignoring distinctions between equivalent observer-moments. First some notation. If $\equiv$ is an equivalence relation on $W*$ and $A \subset W*$, we write $A_{\equiv}$ for the closure of A under $\equiv$ and $A \bmod \equiv$ for A modulo the equivalence relation $\equiv$. Thus $x \in A_{\equiv}$ if and only if there exists $y \in A$ with $y \equiv x$ and $g \in A \bmod \equiv$ if and only if g is an equivalence class modulo $\equiv$ and $g \cap A \neq \emptyset$. We shall write $x_{\equiv}$ rather than $\{x\}_{\equiv}$ in the case of a singleton set. If $z \in W*$, we can extend P_z so that $P_z(A)$ is defined for any $A \subseteq W*$ by setting $P_z(A) = P_z(A \cap R_z)$. This is equivalent to enlarging the reference class R_z to all of $W*$ but giving zero probability to all sets of observer-moments that do not intersect R_z. We can easily define a prior probability distribution, which we will represent by P_z even if this is an abuse of notation, on $W* \bmod \equiv$; just set $P_z(A \bmod \equiv) = P_z(A_{\equiv})$ for $A \subseteq W*$.

What we really want to know is when

$$P_z(A_{\equiv}|(K_z)_{\equiv}) = P_z(A|K_z). \tag{1}$$

Equation 1 is stated in terms of closures but there is an equivalent equation that can be written that relates conditional probabilities of equivalence classes. It is just more convenient to discuss closures than equivalence classes.

The distinction between observer-moments that belong to the same equivalence class is not supposed to matter so we shall assume that we only care about sets A that are closed under $\equiv$. ($A_{\equiv} = A$.)

By definition of conditional probability, equation 1 is equivalent to

$$\frac{P_z(A_{\equiv} \cap (K_z)_{\equiv})}{P_z((K_z)_{\equiv})} = \frac{P_z(A \cap K_z)}{P_z(K_z)}. \tag{2}$$

Assuming $P_z(A \cap K_z) \neq 0$, we know that $P_z(A_{\equiv} \cap (K_z)_{\equiv}) \neq 0$ and we can derive

$$\frac{P_z(A \cap K_z)}{P_z(A \cap (K_z)_{\equiv})} = \frac{P_z(K_z)}{P_z((K_z)_{\equiv})} \tag{3}$$

where we have taken into account that $A_{\equiv} = A$. Equation 3 just says that if we select an element at random from $(K_z)_{\equiv}$, the probability that the element actually belongs to the knowledge set K_z does not depend on whether the chosen

element belongs to the set A. Here the random selection was done using the probability distribution P_z.

Equation 3 will be true for example if for all equivalence classes g in $K_z \mathrm{mod} \equiv$, the probability that a random element of g actually belongs to K_z is the same.

Let us examine some important special cases. Let $x \equiv y$ if and only if $x* = y*$. (So to be equivalent, observer-moments must belong to the same world. And there is a correspondence between equivalence classes of observer-moments and uncentered worlds. So when we ask for the posterior probability of $A = A_{\equiv}$, we are asking for a posterior probability for a set of possible worlds.) Assume if w is a world consistent with K_z (a world that for all z knows might be actual), the probability is k that a random observer-moment $y \in w*$ actually belongs to K_z (for all z knows she could be y if $y \in K_z$) where k is a constant. Then equation 3 and hence equation 1 is satisfied.

This might happen if we have reasonable counting measure $|\quad|$ on observer-moments in a given world (for example the total amount of information represented or the total number of atomic moments) such that for all worlds w that for all z knows might be actual, the count $|w*|$ is the same and the count $|w* \cap K_z|$ is the same (for example it might always equal 1. This is saying that all worlds consistent with z's knowledge have the same number of observer-moments and in every such world there is a constant number of observer-moments such that for all z knows she could be they. Even if this is not exactly true, it might be approximately correct and thus show why we can often ignore strictly anthropic information when reasoning about (bare) possible worlds.

We are also interested in the equivalence relation that considers two observer-moments to be equal if and only if they belong to the same observer in the same world. Assume once again that we have a reasonable way of counting observer-moments such as total amount of information or total number of atomic moments. We also require that for any o and w such that for all z knows she might be a part of o in w, the count of observer-moments in o in w is the same. Finally assume that if for all z knows she might be o in w, the number of observer-moments y in o in w such that for all z knows she might be y does not depend on o and w. This is saying that $|o|$ and $|o \cap K_z|$ is the same for all such o and w. We can then derive the result that we (i.e. z) can safely ignore temporal location information, if we have a question about which world is actual or about our identity but do not really care what time it is.

We can now see why it is often possible to estimate approximately correct posterior probabilities by conditionalizing a chronological prior. It is true enough that almost always if t_2 is a later time than t_1, we will have lost some information between t_1 and t_2 and strictly speaking if information is lost, between t_1 and t_2, it would not be correct to conditionalize our probability distribution at time t_1 on the knowledge acquired in the time between t_1 and t_2. But the information lost might be irrelevant information.

We might at t_1 know the current time is t_1 and at t_2 we do not know that the current time is t_1 but temporal location might be irrelevant. We might at t_2 knows some nonanthropic information E that we did not know at t_1. So we lost the information that we did not E. However even if E might be very relevant

information, whether or not we know E might not be relevant information. In the Doomsday Argument scenario, information about birth rank is very relevant to estimation of the probability of Doom Soon but information about whether one knows one's birth rank is not relevant.

If we do not lose relevant information between t_1 and t_2, then we should be able to conditionalize our probability distribution at t_1 on the knowledge acquired between t_1 and t_2.

Let y and z be two different observer-moments in the same possible world. We would like to show that z can compute her relevant posterior probabilities by starting with a distribution representing y's relevant knowledge and then conditionalizing this distribution on the relevant knowledge that z has and y lacks. We let $\equiv$ represent the equivalence relation such that $x \equiv s$ if and only if x and s belong to the same observer in the same possible world. We are interested in the posterior probability $P_z(A|K_z)$ for some set $A \subseteq W*$ with $A = A_{\equiv}$. We would like to show that this posterior probability estimate can be obtained by conditionalization of y's relevant prior if $(K_z)_{\equiv} \subseteq (K_y)_{\equiv}$. (Thus no relevant information known to y is not known to z.)

The distribution P^y representing y's relevant knowledge is defined by

$$P^y(B) = P^y(B_{\equiv}|(K_y)_{\equiv}) \tag{4}$$

for all $B \subseteq W*$. So we would like to be able to say that

$$P_z(A|K_z) = P^y(A|(K_z)_{\equiv}). \tag{5}$$

But

$$P^y(A|(K_z)_{\equiv}) = \frac{P^y(A \cap (K_z)_{\equiv})}{P^y((K_z)_{\equiv})} \tag{6}$$

and

$$P^y(A \cap (K_z)_{\equiv}) = \frac{P_y(A \cap (K_z)_{\equiv})}{P_y((K_y)_{\equiv})} \tag{7}$$

and

$$P^y((K_z)_{\equiv}) = \frac{P_y((K_z)_{\equiv})}{P_y((K_y)_{\equiv}} \tag{8}$$

where we have used the fact that $(K_z)_{\equiv} \cap (K_y)_{\equiv} = (K_z)_{\equiv}$.

Thus

$$P^y(A|(K_z)_{\equiv}) = \frac{P_y(A \cap (K_z)_{\equiv}}{P_y((K_z)_{\equiv})} \tag{9}$$

.

If we assume that y and z use the same priors and thus $P_y = P_z$, the right hand side of equation 9 is just $P_z(A|(K_z)_{\equiv})$.

Therefore if equation 1 is true of A and z and $\equiv$, then it is permissible for z to obtain her relevant posterior by conditionalizing y's relevant posterior (which z will treat as a prior) on the (relevant) knowledge that z has and y does not.

This proof did not really depend on any chronological relationship between y and z nor do y and z have to belong to the same observer in the same world.

But our assumption of no loss of relevant information is most plausible if z occurs after y and z and y belong to the same observer.

13 Anthropic Causal Decision Theory

So far we have only considered scenarios in which observer-moments have to estimate probabilities or make decisions that do not affect the total number of observer-moments that exist in the actual world. But certainly people can cause other people to die or cause other people to be born. People may not be able to figure out the long range consequences of performing actions that will have the effect of shortening their own or someone else length of life or the long range consequences of performing actions that will cause people to be born. Maybe the long range consequence of my shooting a dictator and his most powerful subordinates might be to increase the total number of observer-moments that will exist in the actual world. Maybe the long range consequence of my choosing to have twenty children is to decrease the total number of people who will ever be born. Even governments may make mistakes when they engage in pronatalist or population control policies domestically or internationally. But even if we cannot know for sure what the consequences are of such actions we can take actions that do affect the total number of observer-moments that will ever live on earth and we can make probabilistic statements about the consequences of such actions. So we do need a theory that will allow us to reason about which decisions we should making when decisions might affect the total number[47] of observer-moments that will ever live.

Scenarios even more troublingly paradoxical than the Doomsday Argument scenario can be constructed if we allow observers to make decisions that affect the total number of observer-moments who exist. We might consider a modified version of the Lazy Adam scenario of Bostrom[Bos02a].

There are two possible worlds Doom Soon and Doom Later[48]. In Doom Soon worlds there is only one observer. In Doom Later worlds there are N observers with N much greater than 1 (e.g. 10^{100}). All observers in both kinds of worlds consist of the same number of observer-moments. There are two kinds of Doom Soon worlds, worlds in which event E happens and worlds in which event E does not happens. There are also two kinds of Doom Later worlds, worlds in which E happens (but in some versions of our scenario there will not actually be any Doom Later worlds with E occurring) and worlds in which E does not happen. The event E might be an event that there are good nonanthropic reason to believe rather improbable, for example throwing a fair coin 100 times and obtaining heads each time or a wounded but otherwise healthy deer wandering into the backyard of the observer with birth rank 1 (call this observer Adam) on a certain specific day without any effort on Adam's part.

At some time before the time at which event E might occur, Adam has a

[47] the total number according to some reasonable counting measure

[48] Actually these are equivalence classes of worlds and later we shall have to distinguish between worlds within the same equivalence class.

decision to make. If he chooses option A, the actual world will be Doom Soon. If he chooses option B and E does occur the actual world will be Doom Soon. If he chooses option B and E does not occur, then the actual world will be Doom Later[49] Whatever actions are involved in choosing A or B do not seem to affect whether E occurs. The decision might be a choice to make or not make an irrevocable resolution to push a button of a cloning machine (a machine that will clone Adam) if E does not occur.

On strictly nonanthropic grounds the probability that E occurs given Adam makes the choice A is very small. His option A does not affect how a fair coin lands or where a deer chooses to visit. Nor should his choice B affect whether or not E occurs. Thus we would conclude that using only nonanthropic information that if Adam chooses B, then there is a very high probability that E does not occur and that the actual world is Doom Later. But this is a nonanthropic probability estimate. Adam might want to revise this nonanthropic estimate to take into account the fact that he is Adam, the observer with birth rank 1. The nonanthropic probability of Doom Later given choice B is made is very high but given that it is very atypical for a random observer in Doom Later to have rank 1 but in Doom Soon there is only one observer, a Doomsday Argument will show that given that choice B is made, Adam should believe that the probability of Doom Soon and hence of E happening is very high. Adam wants E to happen and thus it would seem he would make choice B and thus Adam can be almost certain that E will happen even if it is an event like a fair coin landing heads a hundred times in a row.

If we ignore the issue of whether Adam should choose option B and just assume that Adam is under a compulsion to choose B, a compulsion that is irresistable and that every observer knows that Adam has this compulsion and that the compulsion is irresistable, we are not in a situation at all very different than the standard Doomsday Argument scenario (we might even imagine their is a phase early in all observers' lives when they are totally ignorant of their birth rank). It should not matter the exact mechanism that is responsible for the fact that the actual world is Doom Soon rather than Doom Later or vice versa. It should not matter whether that mechanism involved purely cosmological events or also involved the actions of observers such as Adam. It should not matter that Adam has a strong preference that the actual world be Doom Soon. There is nothing really new here. And nothing changes that much if there is some

[49] A more sophisticated version of this scenario stipulates that the relation between whether E occurs, Adam's decision and whether the actual world is a Doom Soon or Doom Later world is stochastic. Adam can try to force the actual world to be a Doom Later world whenever E does not occur, but he cannot be absolutely sure he will succeed. Adam might correctly believe that if he chooses A, the actual world is most likely to be Doom Soon but he cannot know for sure that it will be Doom Soon.

In Bostrom's Lazy Adam scenario, the decision involves an irrevocable resolution to reproduce if it is discovered that E does not occur and one might wonder if one can make such an irrevocable resolution[Cir04]. (But one might push a button on a computer that will assess whether E occurs and if necessary force Adam to reproduce.) And even if one cannot make a perfectly irrevocable decision, one might be able to make a resolution that is not very likely to be revoked.

stochastic process that determines whether Adam succombs to the compulsion to choose B or actually chooses A. Adam is part of nature. Adam can still calculate nonanthropic and anthropic probabilities for Doom Soon. We find Adam's high probability estimate for Doom Soon and hence for E troubling. But what we are really troubled by is the sharp difference between anthropic and nonanthropic probabilities.

We can also consider a kind of repeated Lazy Adam scenario, which is even more troubling. Possible worlds have S binary digit id numbers. The number of observers who live in a world with id number x is N^{x^+} where x^+ is the number of nonzero digits in x. But we are primarily concerned with Adam, the observer with rank 1. There are for $1 \leq i \leq S$, events E_i that might or might not occur and associated with each event is a decision for Adam to make. Adam can choose either A_i or B_i. Adam must make the choice before he has any chance to observe whether or not E_i occurs. If Adam chooses A_i or if Adam chooses B_i and E_i occurs, then the ith digit of the number of the actual world is 0. Otherwise the ith digit is 1. Adam is under an irresistable compulsion to choose A_i for $i \in C$ and B_i for $i \notin C$ where C is some subset of the set of integers between 1 and S such that the fraction of S that belongs to C is around one-half. The nonanthropic probabilities for the actual world being a world in which E_i occurs are very small for all i and these nonanthropic probabilities are independent. But once Adam takes into account his knowledge of birth rank, he will conclude that if $i \in C$, then almost certainly E_i did occur but if $i \notin C$ almost certainly E_i did not occur. This seems a strange coincidence. But we could have a similar strange coincidence without any connection with any decision-making of Adam's. What if all the E_i were cosmological events and that these were actually rather similar events but only if $i \notin C$ does event E_i not occurring cause the actual world to have digit 1 in the ith dimension and for all the other i, the digit is 0. Adam would still notice the same strange discrepency between the objective improbability of all the E_i and high probability of only those E_i whose nonoccurance would cause the number of observer-moments to be multiplied by N. This is basically just the same discrepency between anthropic and nonanthropic probability that we have seen before arises in the Doomsday Argument scenario. Once we have an anomaly we can often play around with the anomaly and make it arise in a strange pattern of places but there is nothing really new here.

But the issue arises: Should Adam in the original Lazy Adam scenario choose A or B. If the choice of A or B does not really affect objectively (does not affect the nonanthropic probability of) the likelihood of E occurring, perhaps we should want to conclude that it does not matter if Adam chooses A or B so why not choose B. But to make matters more interesting we assume there is a cost associated with chosing B rather than A. A fairly large cost but a cost that Adam is willing to pay if he could cause it to be the case that E is highly likely to occur rather than highly unlikely. We might also make things more interesting by assuming that Adam has to choose A or B before he knows that he is Adam (and we assume that this choice has no effect when made by an observer other than Adam).

I think it is fairly clear that Adam should choose A not B. A good version of causal decision theory should tell us why. The basic idea is when deciding whether to choose A or B Adam should ask himself: If I choose A in the actual world, would I be better off than if I had chosen B and if I chose B in the actual world, would I be better off than if I had chosen A[50]. If the answer to the first question is yes and the second no, then A is to be preferred. If the answers are no and yes, then B is to be preferred. If both answers are no or both answers are yes, then more reasoning needs to be done. We are in a sort of paradoxical situation.

In order to help him to decide what utility he would have if he had made a certain choice, a basic insight that Adam can use is that if in the actual world he is the observer with rank 1, he would still be the observer with rank 1 if he had made a decision different than the decision that he actually made.

Consider now the situation where Adam chooses A (actually chooses A). And E does not occur. His choosing B would not cause him not to have birth rank 1 or cause E to occur. It would just cause Doom Later to be actual rather than Doom Soon and there would be a cost associated with choosing B so B should not be chosen. But note that this argument involved a comparison between an actually possible world in the set W and some hypothetical world, which might not be part of W. In fact Adam might be compelled to choose A by a law of nature but we could still reason counterfactually about the consequences of having made the impossible choice B.

Consider next the situation where Adam chooses B (actually chooses B). And E does occur. His choosing A would save him a cost and not prevent him from having birth rank one or prevent E from occurring. So B should not be chosen.

Certainly if Adam chooses A and E does occur, then he would not be better off if he had chosen B (and E would occur anyway). And if Adam chooses B and E does not occur, then that is not something he wants so he might as well have chosen A.

Thus no matter which world is actual, it is best if Adam chooses A.

The Lazy Adam scenario is only one example of a scenario where making a decision is not so simple and the difference between epistemic and causal decision theory might matter[Joy99]. We need to investigate how we might adjust causal decision theory so that we can handle problems involving decisions that might affect the number of observer-moments in the universe. A sketch of a theory is presented below.

13.1 Sketch of an Anthropic Causal Decision Theory

We consider an observer-moment $z \in W*$ who is making a decision D. We assume that the decision is a decision between a finite number S of alternatives. We call the alteratives D_i for $1 \leq i \leq |S|$. The decision D might be an actual

[50]Of course, Adam does not know which world is actual so we might have to talk about expected utilities and about whether Adam could on average expect to be better off rather than whether he is actually better off or not.

decision that z is faced with or some hypothetical decision problem might be involved. Since we do not want the mere fact that she is faced with making the decision D to give z any knowledge about her location within the set K_z, we must assume that all observer-moments in K_z are faced with a similar decision problem. In fact, we can assume that all $y \in K_z$ are faced with decision D.

We want to analyze the effect of z's decision on her utility. Unfortunately in general her utility might depend not just on her own decision but on the decisions of other $y \in K_z$. That complicates the analysis. So we shall only treat the simple case where there is at most one element of K_z in each $w \in W$ or we assume that the effects on utility of different decisions by different members of $K_z \cap w$ are independent. Thus even if x and y are both in world w and both in K_z, the difference in utility x would obtain by choosing D_i rather than D_j for $1 \leq i, j \leq S$ is independent of the choice made by $y \neq x$.

Under these assumptions, we can define a function $U : K_z \times S \to \mathcal{R}$. Here $U(y, i)$ is supposed to represent the utility for observer-moment y in choosing D_i if it were faced with decision problem D[51]. In order to compute $U(y, i)$, we have to determine which possible world would result if y in world w were faced with D and chose D_i. We might choose the world closest (in some sense) to w among those worlds in which y chooses D_i. That world might be w itself or it might be some other world and we can then determine the utility of y in w or in the other world. Or there might not be one single definite world that would have resulted if y chose D_i; if y chose D_i, that would set in motion a stochastic process that does not occur in w and there is not enough information in w to tell us what the result of the stochastic process would be. So we might be dealing with a probability distribution over counterfactual possible worlds. But in that case we could compute the utility of y in these worlds and then compute an expected value. In any case, we could compute a value for $U(y, i)$. All we need is that we can determine the utility that would have been obtained by y if she were faced with D and chose D_i. It is not necessary that we use a closest possible world approach to interpreting counterfactuals.

Now what z has to do is fairly clear. She has to evaluate the expected utility $\sum_{y \in K_z} P_z(y|K_z)U(y, i)$ for all i and chose the i that maximizes this expected utility. The problem lies with the $P_z(y|K_z)$. If Adam has to decide between A and B at a time when he does not know who he is, then if y is an observer-moment that is not part of Adam, how can Adam give a probability value for $P_z(y|K_z)$? Whether Doom Later exists depends on how Adam decides and thus whether y exists can depend on a decision of Adam's. In fact it might depend on the very decision problem D that Adam (and every other observer-moment in the same knowledge state as he) is in the process of trying to solve.

Adam might try to make a psychological model of how he and other observer-moments in the same knowledge state would most probably make their decisions[52]. Fortunately, in the Lazy Adam scenario, the exact numbers do not

[51] It is assumed that all $x \in K_y$ knows the value of $U(y, i)$; they know the utility they would be obtain by choosing D_i if they were observer-moment y.

[52] Actually, as we shall see below, Adam needs a probability distribution over S. It might matter exactly how likely each D_i is to be the choice that is actually made. But for simplicity,

matter. No matter what probabilities we use for our $P_z(y|K_z)$, we know Adam would be better off or at least as well off choosing A. In general, there is a potential difficulty because there might be problems where if Adam thinks it likely he will make choice A, then he should make choice B and vice versa[Ega07]. Thus Adam cannot make a choice that he will not later regret.

It is not entirely clear what we should do if we wish to avoid making a decision that we shall regret, but in a context where anthropic reasoning is not a concern, Arntzenius has suggested a possible solution[Arn08]. Inspired by his work, we might suggest the following partial solution.

We start with a theory of psychology which tells us how observer-moments in K_z would decide when faced with D. This would allow us to calculate the P_z we need. Then we might figure out an optimal decision. However, if there is a difference between our initial psychological theory about how likely each D_i is to be chosen and our theory as to which D_i is best to chose, then there is a problem. Thus if our theory is that all observer-moments would choose D_i and we use that theory to obtain a P_z and use that P_z when computing expected utilities and then determine that D_i is really not optimal, we lack stability and there is a problem.

Another possible problem is that there might be $i \neq j$ such that if we start with a psychological theory that says that observer-moments in K_z would select D_i, we will discover that D_i is optimal but if we start with a theory that observer-moments will prefer D_j, we will discover that D_j is optimal. So there might be multiple stable solutions. We will not discuss how to choose in that case. Perhaps we know something about the psychology of observer-moments other than their rationality and this knowledge might help to choose between the different stable solutions. One solution might be simpler or seem more natural.

Even if we do not have a theory for the multiple stable solution case, we have a possible remedy if there are no stable solutions. We can expand the space of possible solutions. Instead of choosing a D_i, we choose a probability distribution over the D_i. We start with a psychological theory that gives a probability distribution π over the D_i.[53] Based on that theory we compute a P_z and then we might discover that for any i such that $\pi(D_i) \neq 0$, D_i is as good a choice as any other choice and thus we cannot improve on the probabalistic choice that select each D_i with probability $\pi(D_i)$[54]. Again it is possible that there are several mixed strategies that are stable.

14 Conclusion

We have seen that a simple and natural theory of anthropic reasoning can be constructed based on Bostrom's Strong Self-Sampling Assumption. In the fi-

we assume at first that the psychological theory picks out a unique D_i.

[53] So the theory say we should a randomizing device to help us make our choice.

[54] And that does not mean that it would necessarily be acceptable to make a definite choice of one of the D_i with $\pi(D_i) \neq 0$ because that choice might not be stable. If we used the psychological theory that D_i would be chosen to help compute a P_z, we would eventually find out that observer-moments have a better option available than D_i.

nite case, there are good Dutch Book arguments as well as relative frequency arguments in favor of using maximal reference classes and giving every atomic observer-moment in the same possible world equal prior probability. If we find unrealistic the notion of atomic moment we can just realize that the measure given to a moment is proportional to the logarithm of the number of distinct subjective psychological states it can represent.

This theory of anthropic reasoning is simple and does not require we use the Self-Indication assumption. In fact some of the same arguments that advocates of the SIA could use against our theory of anthropic reasoning could be used against the SIA as it is usually made precise and in favor of a strong version of the SIA. Furthermore some of the ratioale for the SIA fails to take into account that each observer-moment is making its own decisions independently. So it does not matter if it would be better if everyone applied the SIA rather than no one did. What matters is holding constant what other observer-moments do whether it is desirable for an individual moment to apply the SIA.

A natural theory of anthropic decision making, which is really a kind of anthropic causual decision theory can be defined which enables us to make intuitively natural decisions in scenarios like the Lazy Adam scenario. The basic idea is that if an observer-moment had made a decision other than the decision she actually made she would still be the same observer-moment.

Our theory seems to work reasonably well in the finite case even if we have to acknowledge that the conclusion of the Doomsday Argument, an argument that our theory says is valid, is rather counterintuitive. Any theory of anthropic reasoning that is sufficiently general and non-ad hoc will have some counterintuitive consequences.

Our theory does not really tell us very much about how to define a prior probability distribution in the infinite case; it just says represent the infinite as the limit of the finite but does not really say how. This is an crucial lacuna. The actual world might be a world with an infinity of (atomic) observer-moments and it might be very true that if we try to reason about the actual world as if it contained only a finite number of relevant observer-moments, we will make some very wrong probability estimates. We might not be able to ignore the infinite size of the universe.

Another problem is that we are depending on the existence of a reliable and well-justified nonanthropic prior probability that can be used by all observer-moments in all worlds. The very idea of separation of the nonanthropic and anthropic part of an estimation of probabilities might be mistaken. It might not be the simplest procedure. It might be too difficult to come up with a natural prior that can be used by all observer-moments. When we know very little (we only can take into account what all observer-moments know when creating a nonanthropic prior), we might not provknow enough to derive a meaningful prior. It might be the case that we should not first try to obtain a simple nonanthropic prior and then apply a simple theory of anthropic reasoning. Instead we might from the start have to acknowledge that the anthropic and nonanthropic might be mixed up and might have to be generated as a unit. Maybe all we really want is the simplest probabilistic explanation of the data we actu-

ally have[Hut04, Hut10]. All a theory really needs to do is explain the data we actually have and make (probabilistic) predictions of the data we might acquire if we perform certain experiments or follow certain protocols. It is possible that the simplest way of accomplishing that task and the preferred way of achieving that task is not to first create a nonanthropic prior and then to conditionalize on the anthropic information we do have.

Another topic that needs investigating is constructing hyperpriors. We have discussed how we might learn from experience which anthropic reasoning theory is to be preferred and that would includes learning from experience how to handle infinities but that requires we have some way of generating reasonable hyperpriors on proposed theories of anthropic reasoning. We need to analyze how that might be done in some precise, non-ad hoc way rather than just say that we prefer simple and natural theories. We might have to deal with a range of reasonable hyperpriors but that might be the best we can do.

References

[AEH04] Frank Arntzenius, Adam Elga, , and John Hawthorne. Bayesianism, infinite decisions, and binding. *Mind*, 113(450):251–283, 2004.

[Arn02] Frank Arntzenius. Reflections on Sleeping Beauty. *Analysis*, 62(1):53–62, 2002.

[Arn08] Frank Arntzenius. No regrets: Edith Piaf revamps decision theory. *Erkentnis*, 68(2):277 – 297, 2008.

[BC03] Nick Bostrom and Milan Cirkovic. The doomsday argument and the self-indication assumption:reply to Olum. *Philosophical Quarterly*, 53(210):83–91, 2003.

[BL06] Darren Bradley and Hannes Leitgeb. When betting odds and credences come apart:more worries for Dutch Book arguments. *Analysis*, 66(2):119–127, 2006.

[Bos02a] Nick Bostrom. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York, 2002.

[Bos02b] Nick Bostrom. Self-locating belief in big world: Cosmology's missing link with observation. *Journal of Philosophy*, 99(12), 2002.

[Bos07] Nick Bostrom. Sleeping Beauty: A synthesis of views. *Synthese*, 157(1), 2007.

[Bri10] Rachael Briggs. Putting a value on beauty. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology Volume 3*. Oxford, 2010.

[Car83] Brandon Carter. The anthropic principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society of London A*, 310:347–363, 1983.

[Cir04] Milan Cirkovic. Agencies, capacities and anthropic self-selection. *Philosophical Writings*, 27, 2004.

[CL00] Bradley P. Carlin and Thomas A. Louis. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman, 2000.

[Coz07] Mikal Cozic. Imaging and Sleeping Beauty: A case for double halfers. In *Proceeding of the eleventh conference on theoretical aspects of rationality and knowledge*, 2007.

[Die92] Dennis Diecks. Doomsday or: The dangers of statistics. *Philosophical Quarterly*, 42:78–84, 1992.

[DP08] Kai Draper and Joel Pust. Diachronic Dutch Books and Sleeping Beauty. *Synthese*, 164(2), 2008.

[Ega07] Andy Egan. Some counterexampls to causal decision theory. *Philosophical Review*, 116(1), 2007.

[Elg00] Adam Elga. Self-locating belief and the sleeping Beauty problem. *Analysis*, 60(2):143–147, 2000.

[Elg04] Adam Elga. Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, 69(2):383–396, 2004.

[Gil89] Margaret Gilbert. *On Social Facts*. Routledge, 1989.

[Gri84] Patrick Grim. There is no set of all truths. *Analysis*, 44(4):206–208, 1984.

[Hal04] Joseph Halpern. Sleeping Beauty reconsidered: Conditioning and reflection in asychronous systems. In *Twentieth Conference on Uncertainty in AI*, 2004.

[Har55] John C. Harsanyi. Cardinal welfare, individualistic ethics and interpersonal comparisons of utility. *The Journal of Political Economy*, 63(4):309–321, 1955.

[Hit04] Christopher Hitchcock. Beauty and the bets. *Synthese*, 139(3):405–420, 2004.

[HM05] Dien Ho and Bradley Monton. Anthropic reasoning does not conflict with observation. *Analysis*, 65(1):42–45, 2005.

[Hut04] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Springer, 2004.

[Hut10] Marcus Hutter. A complete theory of everything will be subjective. *Algorithms*, 3(4):329–350, 2010.

[Jay03] E. T. Jaynes. *Probability Theory, the Logic of Science*. Cambridge University Press, 2003.

[Joy99] James Joyce. *The Foundations of Causal Decision Theory*. Cambridge, 1999.

[Kad04] J. B. Kadane. Stopping to reflect. *Journal of Philosophy*, 101(6), 2004.

[Kap95] David Kaplan. A problem in possible world semantics. In Walter Sinnott-Armstrong, Diana Raffman, and Nicholas Asher, editors, *Modality, Morality, and Belief: Essays in Honor of Ruth Barcan Marcus*. Cambridge, 1995.

[KM05] Brian Kierland and Bradley Monton. Minimizing inaccuracy for self-locating belief. *Philosophy and Phenomenological Research*, 70(2):384–395, 2005.

[Les92] J. Leslie. Doomsday revisited. *Philosophical Quarterly*, 42(166):85–87, 1992.

[Lew79] David Lewis. Attitudes de dicto and de se. *Philosophical Review*, 88:513–543, 1979.

[Lew01] David Lewis. Sleeping Beauty: A reply to Elga. *Analysis*, 61(3):171–176, 2001.

[Mea08a] Chris Meacham. Sleeping Beauty and the dynamics of de se belief. *Philosophical Studies*, 138(208), 2008.

[Mea08b] Christopher Meachum. Sleeping Beauty and the dynamics of de se belief. *Philosophical Studies*, 2008.

[Mos09] Sarah Moss. Updating as communication. In *Formal Epistemology Workshop*, June 2009.

[Nea06] Radford Neal. Puzzles of anthropic reasoning resolved using full non-indexical conditioning, 2006. arXiv:math.ST/ 0608592.

[Olu02] Ken Olum. The doomsday argument and the number of possible observers. *Philosophical Quarterly*, 52(207):164–184, 2002.

[Olu04] Ken Olum. Conflict between anthropic reasoning and observation. *Analysis*, 64(1):1–8, 2004.

[Pag07] Don Page. Observation selection effects in quantum cosmology, 2007. arXiv:hep-th/0712.2240.

[Pusar] Joel Pust. Conditionalization and essentially indexical credence. *Journal of Philosophy*, to appear.

[Sol64] Ray Solomonoff. A formal theory of inductive inference, parts i and ii. *Inform. Contr.*, 7:1–22 and 224–254, 1964.

[Sus06] Leonard Susskind. *The Cosmic Landscape: String Theory and the Illusion of Intelligent Design*. Back Bay Books, 2006.

[Tit07] Micheal Titelbaum. *Quitting Certainties: A Doxastic Modeling Framework*. PhD thesis, Department of Philosopohy, University of California, Berkeley, 2007.

[Vil95] Andrei Vilenkin. Predictions from quantum cosmology. *Physical Review Letters*, 74:846–849, 1995.

[vNM44] John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton, 1944.

[Whi10] Roger White. Evidential symmetry and mushy credence. In Tamar Szabo Gendler and John Hawthorne, editors, *Oxford Studies in Epistemology Volume 3*. Oxford, 2010.